The accessible chromatin landscape of the human genome

Thurman et al. 2012

Supplementary Information

Supplementary Datasets

Supplementary Figures

Item	*Descriptive title*
Supplementary Fig. 1	DNaseI density tracks for the 125 cell types analysed
Supplementary Fig. 2	Additional detail for Supplementary Fig. 1, and enhancers
Supplementary Fig. 3	Accessible chromatin peaks overlapping microRNA promoters
Supplementary Fig. 4	DHSs in repetitive elements and a miRNA promoter
Supplementary Fig. 5	Degrees of cell-type-specificity of DHSs in four repeat classes
Supplementary Fig. 6	Quantifying transcription factor impact on chromatin accessibility
Supplementary Fig. 7	Transcription factor occupancies within accessible chromatin
Supplementary Fig. 8	DNaseI and H3K4me3 patterns around promoters in 56 cell types
Supplementary Fig. 9	Overlaps between novel promoters, CAGE clusters, and ESTs
Supplementary Fig. 10	Additional examples of novel promoters identified in K562 cells
Supplementary Fig. 11	Further examples of association between methylation and accessibility
Supplementary Fig. 12	Genome-wide Influence of methylation on chromatin accessibility
Supplementary Fig. 13	Cell-type-specific enhancers at the IFNG locus
Supplementary Fig. 14	Interaction and GO class enrichments via signal-vector correlation
Supplementary Fig. 15	Statistical significance of co-occurrences of motif families
Supplementary Fig. 16	Examples of stereotyped DNaseI patterns across cell lines
Supplementary Fig. 17	Top-ranked matches of stereotyped DNaseI patterns across cell lines
Supplementary Fig. 18	Using a self-organizing map to cluster DHSs by cross-cell-type pattern
Supplementary Fig. 19	Colour-coded key to the cell types in Supplementary Fig. 18
Supplementary Fig. 20	Instance counts of patterns discovered by the SOM (Supp. Fig. 18)

Supplementary Tables

Item	*Descriptive title*
Supplementary Table 1	The 125 cell types analysed, and the sources of their DNaseI data
Supplementary Table 2	Repeat-Masked elements prolifically overlapping DHSs
Supplementary Table 3	Enhancer activity of DHSs overlapping transposable elements
Supplementary Table 4	List of 1046 known regulatory elements, with references
Supplementary Table 5	Mapping of TRANSFAC motif models to gene names
Supplementary Table 6	Merging of DHSs from 79 cell types into 32 categories
Supplementary Table 7	Promoter/distal DHS pairs with correlation ≥ 0.7
Supplementary Table 8	Gene sets and search terms for GO analysis of connected DHSs
Supplementary Table 9	Groupings of TRANSFAC motifs into families and classes
Supplementary Table 10	Replicate data quality and reproducibility

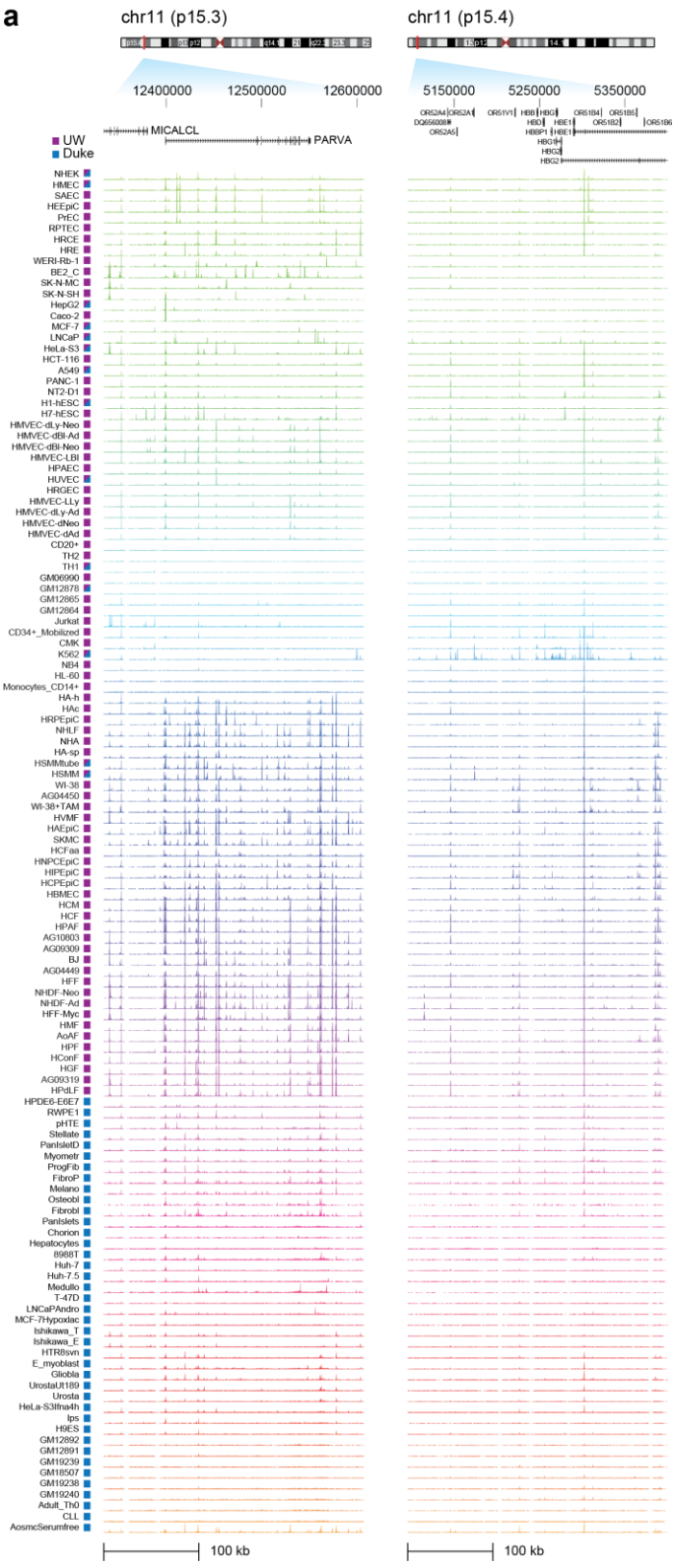
Supplementary Methods

Section	*Title*	*Main text figures*	*Supp. figures*	*Supp. tables*
1.1	DNaseI and histone modification protocols	1a	1, 2	1, 10
1.2	DHS Master List and its annotation	1b,c	–	–
1.3	miRNAs	1c	3	–
1.4	Analysis of Repeat-Masked DHSs	1c	4, 5	2, 3
2	Determining relationships between sequence motifs and chromatin accessibility	2	6, 7	–
3	Promoter DHS identification scheme	3	8, 9, 10	1
4.1	RNA expression	4b-e	11b	–
4.2	RRBS genome-wide methylation profiling	4a-e	11, 12	5
5.1	Connectivity between promoter DHSs and distal DHSs	5a-b	13, 14a	6, 7
5.2	Analysis of 5C data	5a	14b	–
5.3	Gene ontology analysis of DHSs	–	14d	8
5.4	Analysis of sequence motif pairs co-occurring in promoters and connected DHSs	5c	15	9
6.1	DNaseI pattern matching	–	16-18	–
6.2	Self-organizing map	–	19-21	–
7	Measurement of nucleotide heterozygosity and estimation of mutation rate	7	–	–

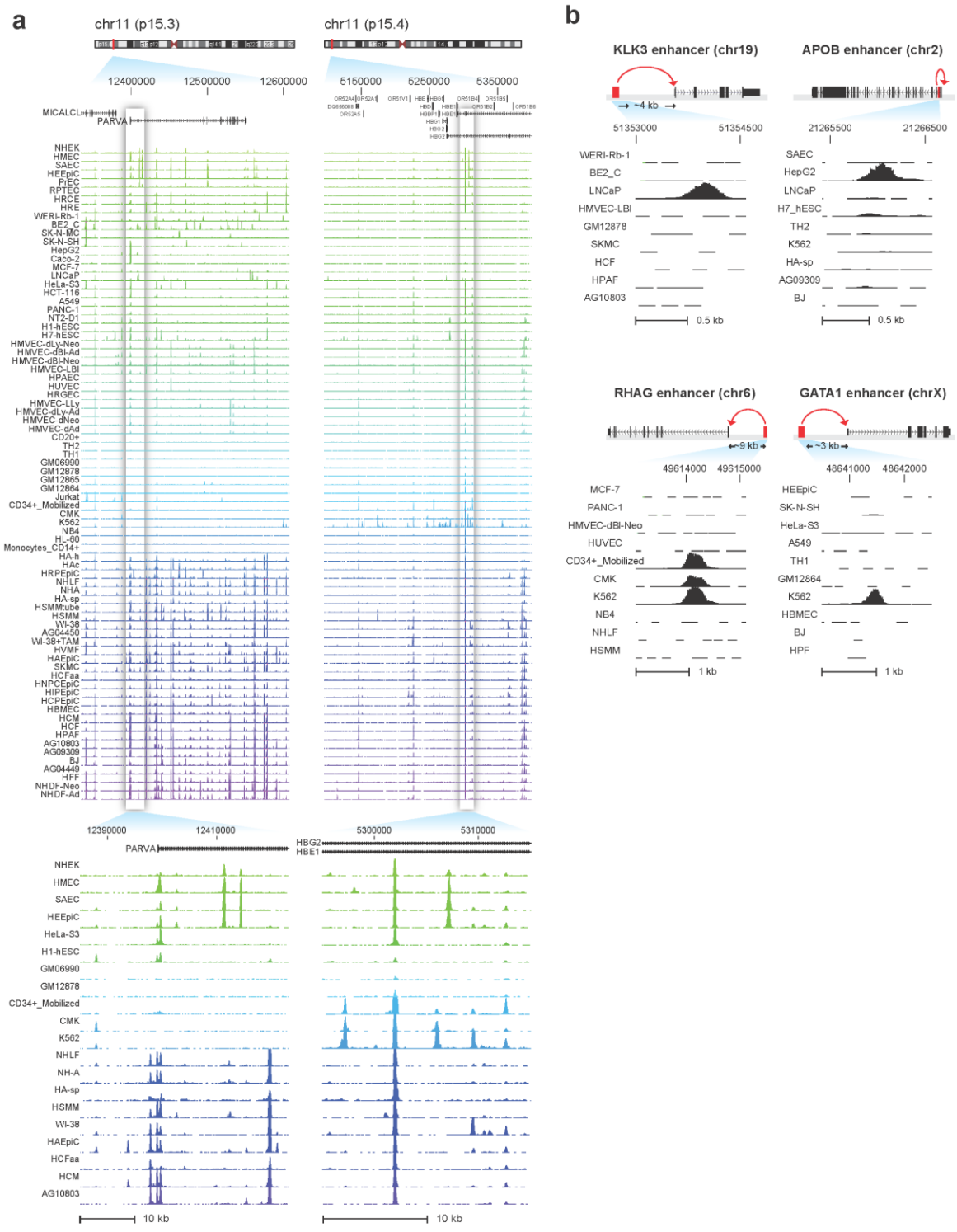
Supplementary References

Supplementary Datasets

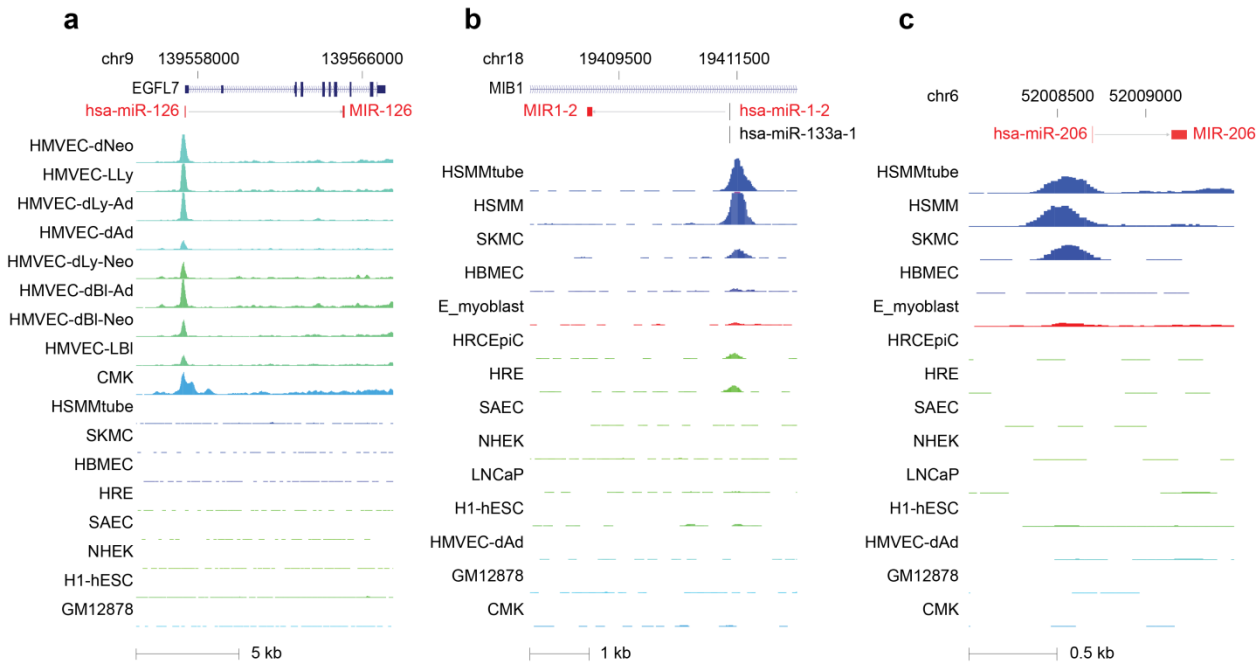
Supplementary files too large to include in this supplement are being made available via the ftp server at ebi.ac.uk which contains an organized file structure with the ENCODE data. Analysis datasets are located in ftp://ftp-private.ebi.ac.uk/ (Login:encode-box-01 Passwd: enc*deDOWN) in the directories under byDataType. Links to such files appear directly in the relevant section of the Supplementary Methods below.



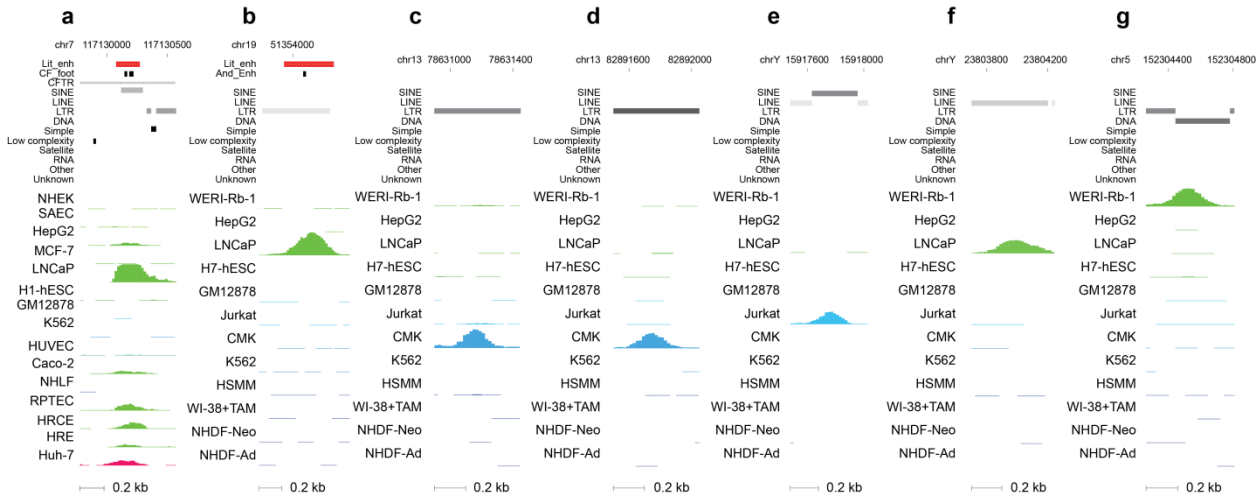
Supplementary Figure 1. (a), Density of DNaseI cleavage sites for all 125 cell types, shown for two example ~350kb regions on chr11. Colour-coded squares next to the cell-type names indicate the origin for the data-set; both centers (UW and Duke) produced data for 14 of the cell types.



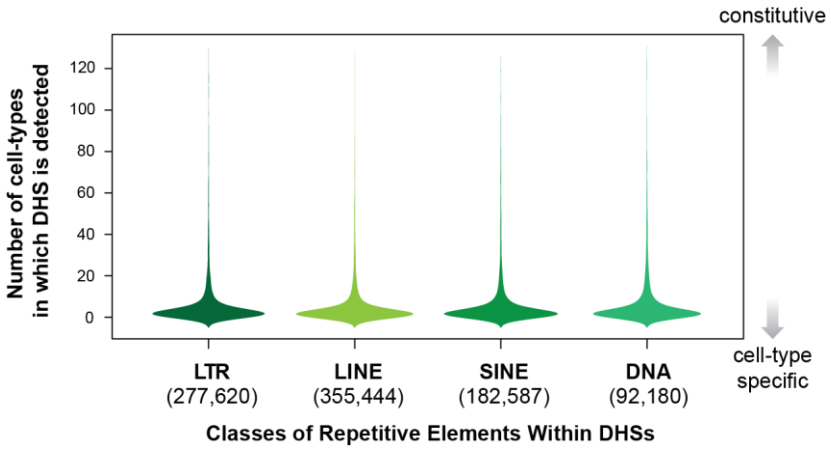
Supplementary Figure 2. (a) This is a subset of the data shown in **Supplementary Fig. 1**, providing greater visual detail. Data for 77 of the 125 cell types is shown. (b) Examples of known cell-selective enhancers. Shown above each set of DNaseI data are schematics showing enhancer location (red) relative to the gene it controls.



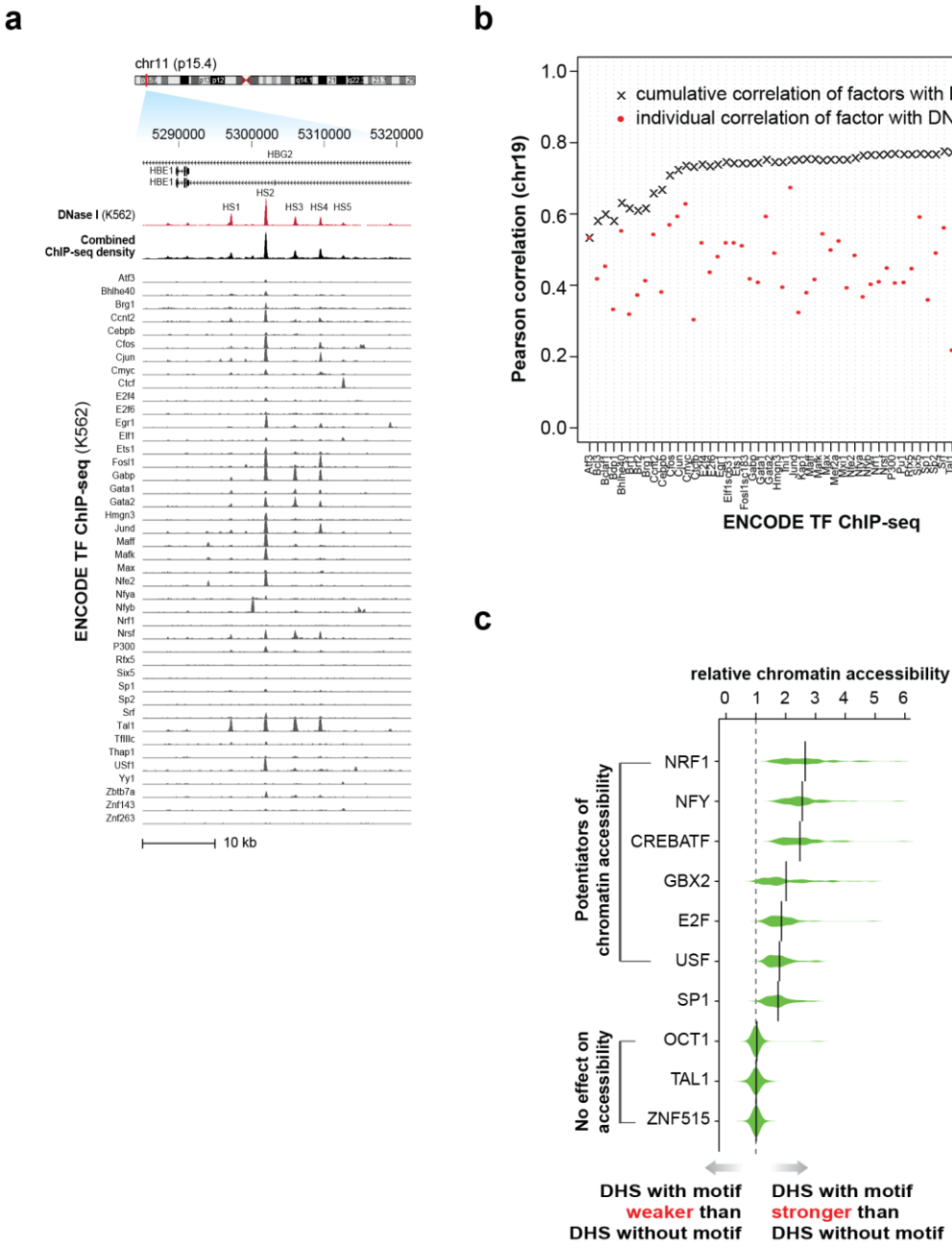
Supplementary Figure 3. Three examples of DHSs overlapping microRNA promoters. Peaks are usually observed in cell types consistent with known function of the microRNA. Panel (a) shows DNaseI signal at the promoter for MIR126. MIR126 is intronic, part of the transcript of the EGFL7 gene. MIR126 has a DHS at the promoter in several endothelial cell lines, consistent with its known function[1]. Panel (b) shows chromatin accessibility at the promoter for MIR1-2. The transcript is antisense of the MB1 gene. DHSs can be seen in muscle cell lines. Panel (c) shows a DHS at a potential promoter site in the muscle cell types HSMM, HSMMtube, SKMC, and myoblast. MIR1-2 and MIR206 are known to be involved in muscle function[2].



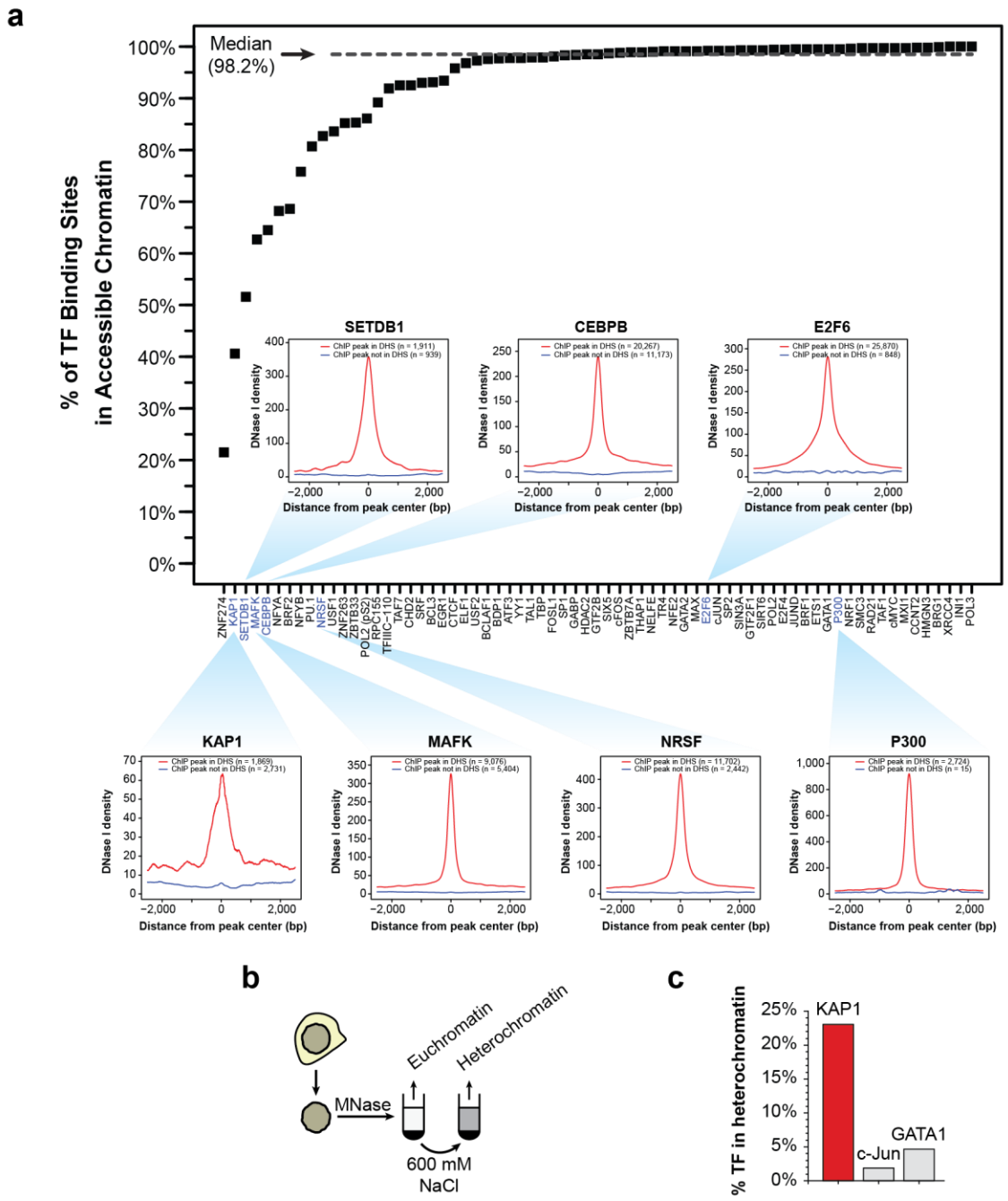
Supplementary Figure 4. Examples of DHSs in repetitive elements and an miRNA promoter. Panels (a) and (b) show data for two well-characterized enhancers which lie in repeat-masked sequence. A CFTR enhancer[3] is shown in panel (a). A red bar marks the position of the literature enhancer which largely overlaps a SINE element. In vitro footprints observed at the enhancer are shown below the red bar in black. The enhancer has been previously reported in Caco-2 and Huh7 cells. We observe a strong signal in LNCaP also. The PSA enhancer of the KLK2 gene[4] shown in panel (b) largely overlaps an LTR element. A red bar marks the known site and a black bar below marks the observed in vitro footprint. A strong DHS is observed in the expected cell type, LNCaP, but not in other cell types. Panels (c)-(g) are examples of DHSs primarily overlapping LTR, SINE, LINE, and DNA elements.



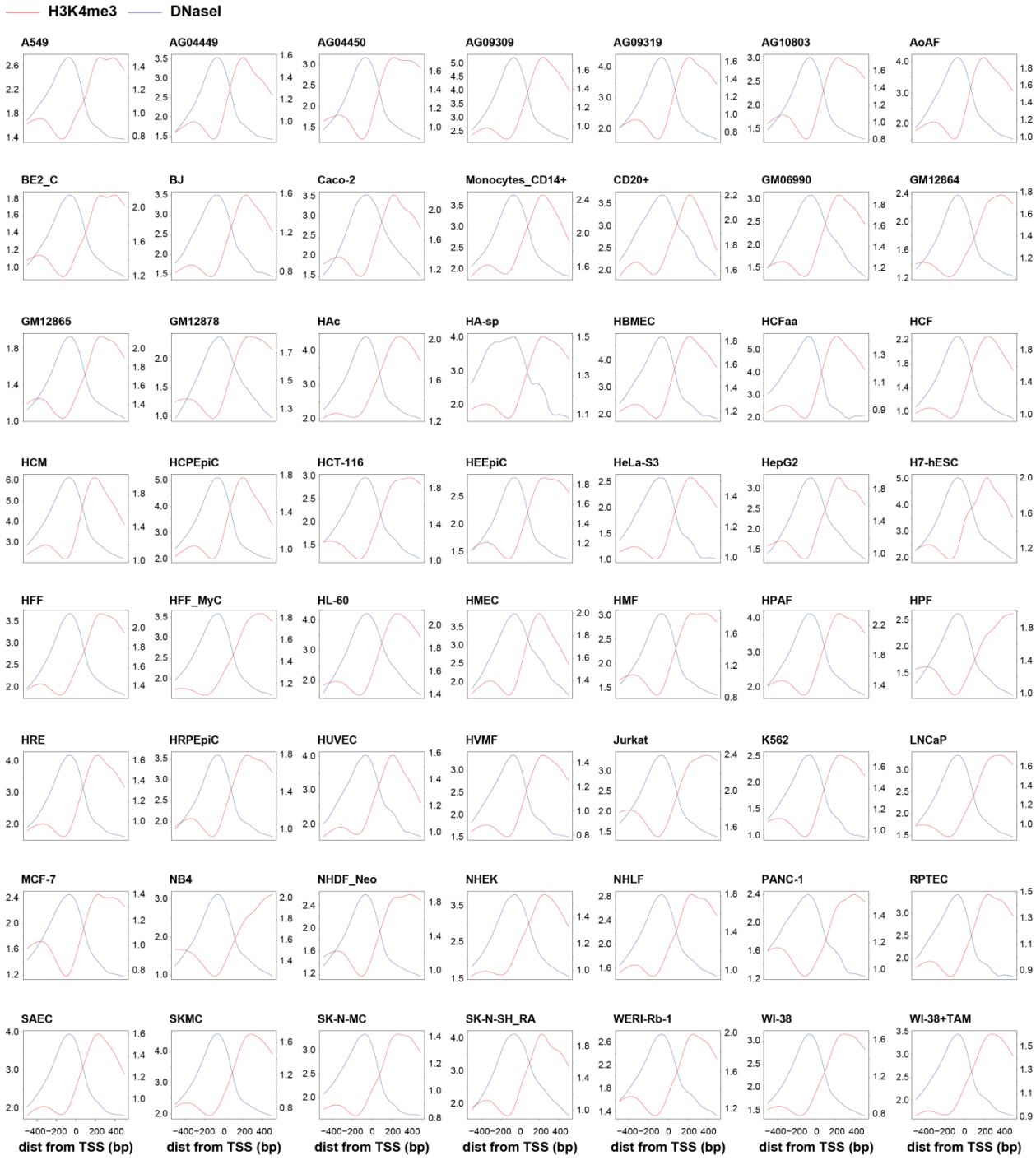
Supplementary Figure 5. Number of cell-types per DHS overlapping four categories of repeat classes. For each master list peak we count the number of cell-types whose peaks overlap at that position, giving a cell-type number per master list peak. The plots show the distribution of these cell-type numbers for DHS overlapping various classes of repeats (RepeatMasker track downloaded from UCSC genome browser). The number below each category is the number of DHS overlapping the repeat class. Average cell-type numbers for each class are: LTR (6.0); LINE (5.3); SINE (5.9); DNA (6.9). This plot was made using the R function "beanplot" from the "beanplot" package.



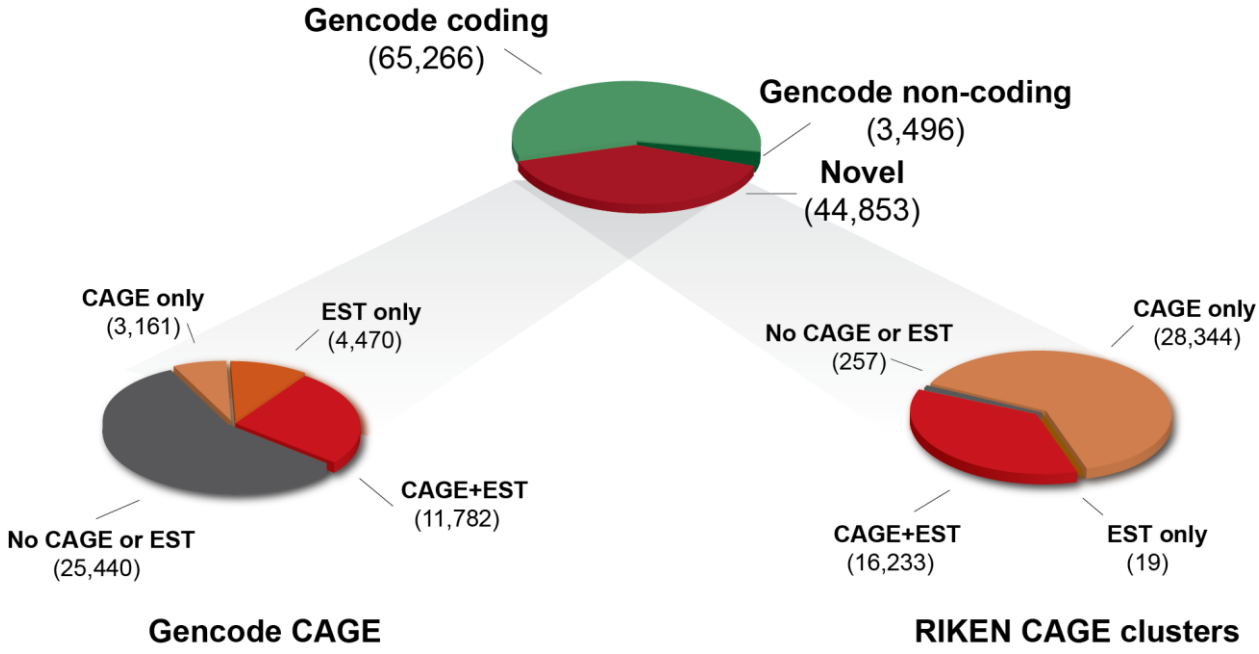
Supplementary Figure 6. Quantifying the impact of transcription factors on chromatin accessibility. (a) As in Fig. 2a, DNaseI tag density is shown in red, followed by normalized ChIP-seq tag density for each of 42 ENCODE ChIP-seq experiments from K562 cells, with a cumulative sum of the individual tag density tracks shown immediately below the K562 DNaseI data; this plot shows a 35 kb region encompassing the beta-globin LCR on Chr11. (b) Additive correlation (y-axis) of ChIP-seq with DNaseI across Chr19 with increasing numbers of TFs. TFs are ordered alphabetically (x-axis). Correlation values for individual factors are shown in red. (c) Relative chromatin accessibility (x-axis) measured as the mean intensity of DHSs containing the indicated motif (y-axis), divided by the mean intensity of all DHSs (using 84 UW DNaseI datasets). Green density plots indicate the distribution of measurements obtained individually across all cell types; values >1 indicate presence of the motif has an average positive effect on chromatin accessibility.



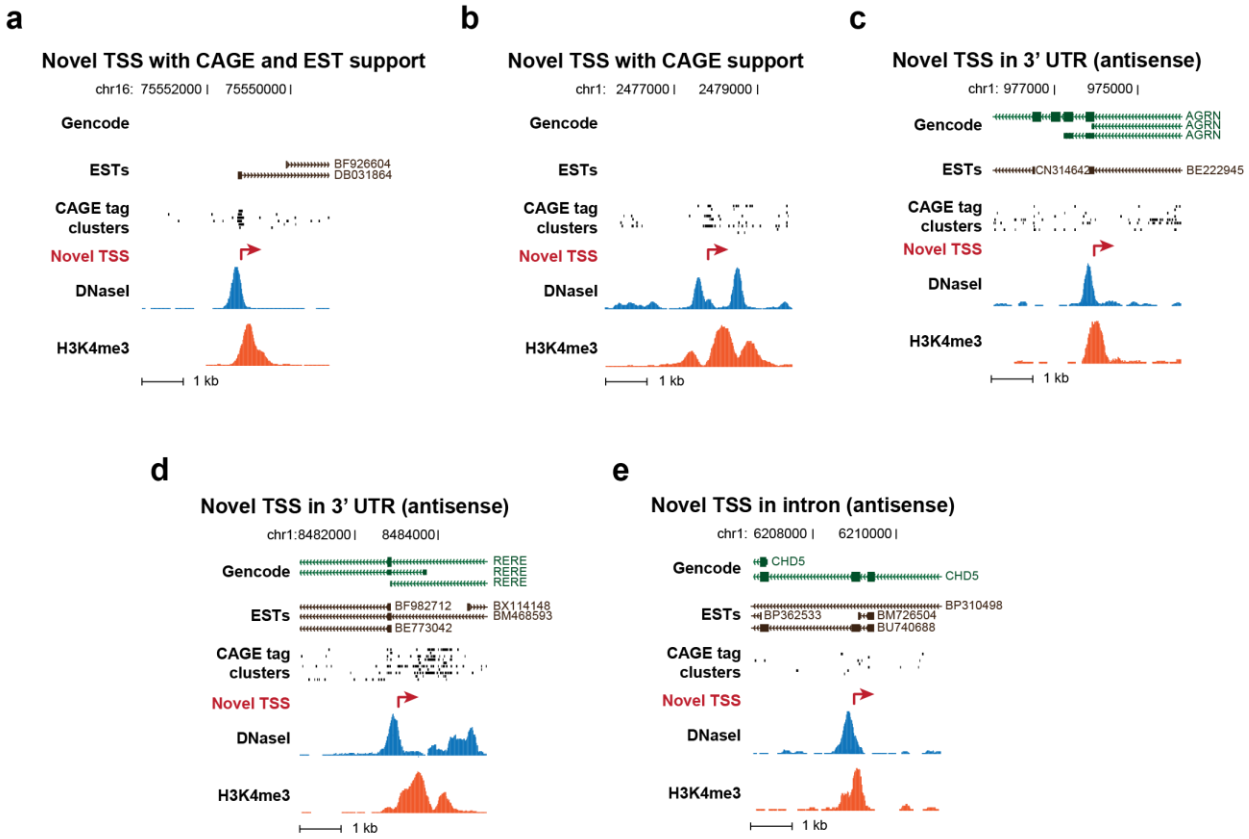
Supplementary Figure 7. The occupancies of different transcription factors within accessible chromatin. (a) The percentage of transcription factor binding sites within accessible chromatin was calculated for each factor. Accessible chromatin was identified using unthresholded hotspot calls on K562 DNaseI deep-seq data. Transcription factor binding sites were identified in K562 cells using ChIP-seq. Inserts show the aggregate DNaseI density profile (±2.5kb of ChIP-seq peak) at sites for six different transcription factors that are within (red) and outside (blue) of accessible chromatin. See Supplementary Methods, section 2.3, below. (b) Biochemical isolation of dense heterochromatin. (c) Proportion of chromatin-bound protein contained within heterochromatin was measured using targeted mass spectrometry for KAP1, c-Jun and GATA1. Note that nearly 25% of nuclear KAP1 localizes to highly compacted heterochromatin, vs. <5% for c-Jun and GATA1.



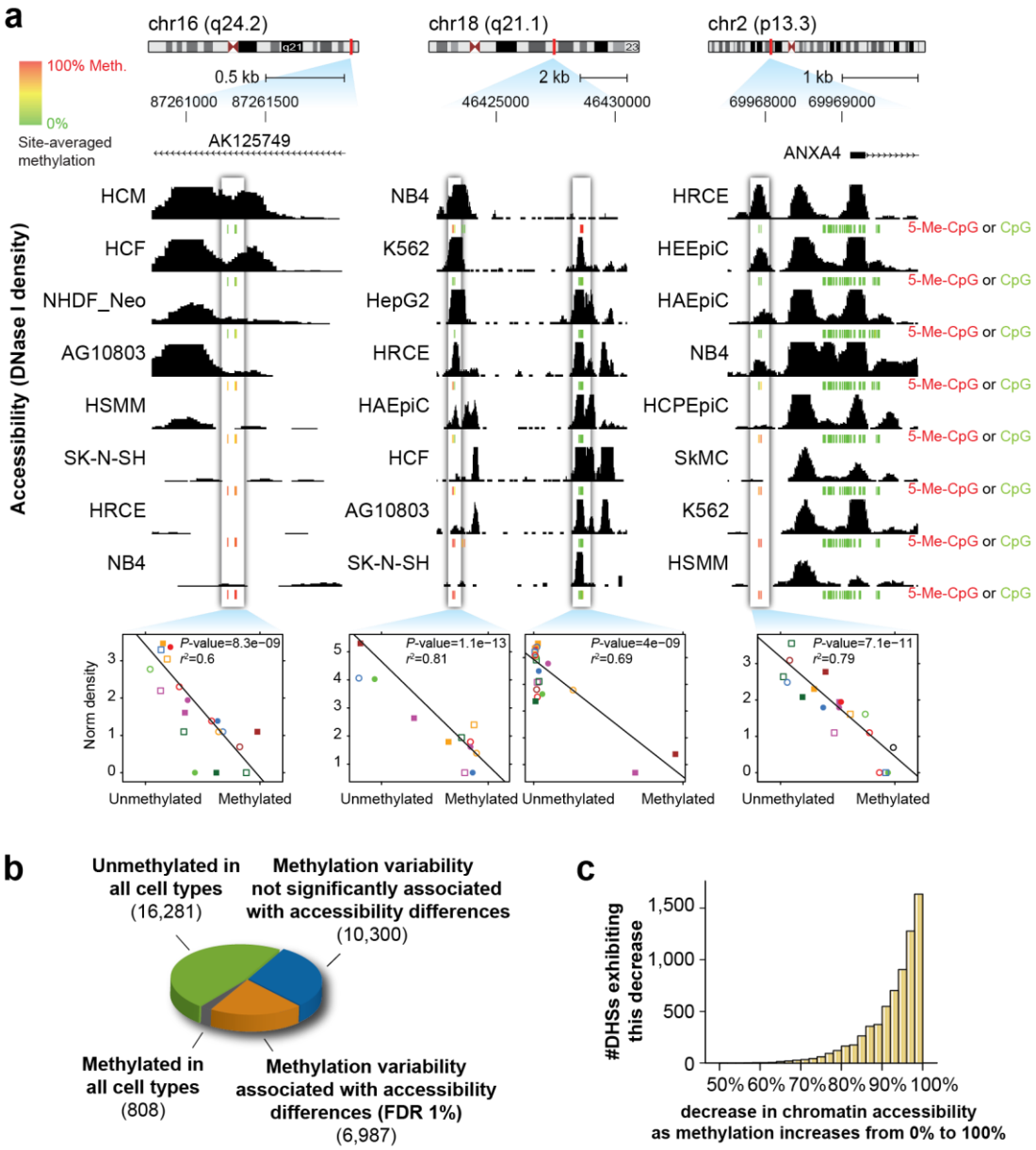
Supplementary Figure 8. This is the same as Fig. 3c, broken out for each of the 56 cell-types for which we have both DNaseI and H3K4me3 data, showing the stereotypical pattern of DNaseI and H3K4me3 around annotated promoters. Tag density for H3K4me3 (red) and log tag density for DNaseI (blue), averaged and centered across 10,000 randomly-selected Gencode v7 TSSs, oriented with respect to the transcription direction (gene body to the right). The x-axis is the distance in bp from the TSS. Left y-axis scale is for DNaseI; right y-axis scale is for H3K4me3.



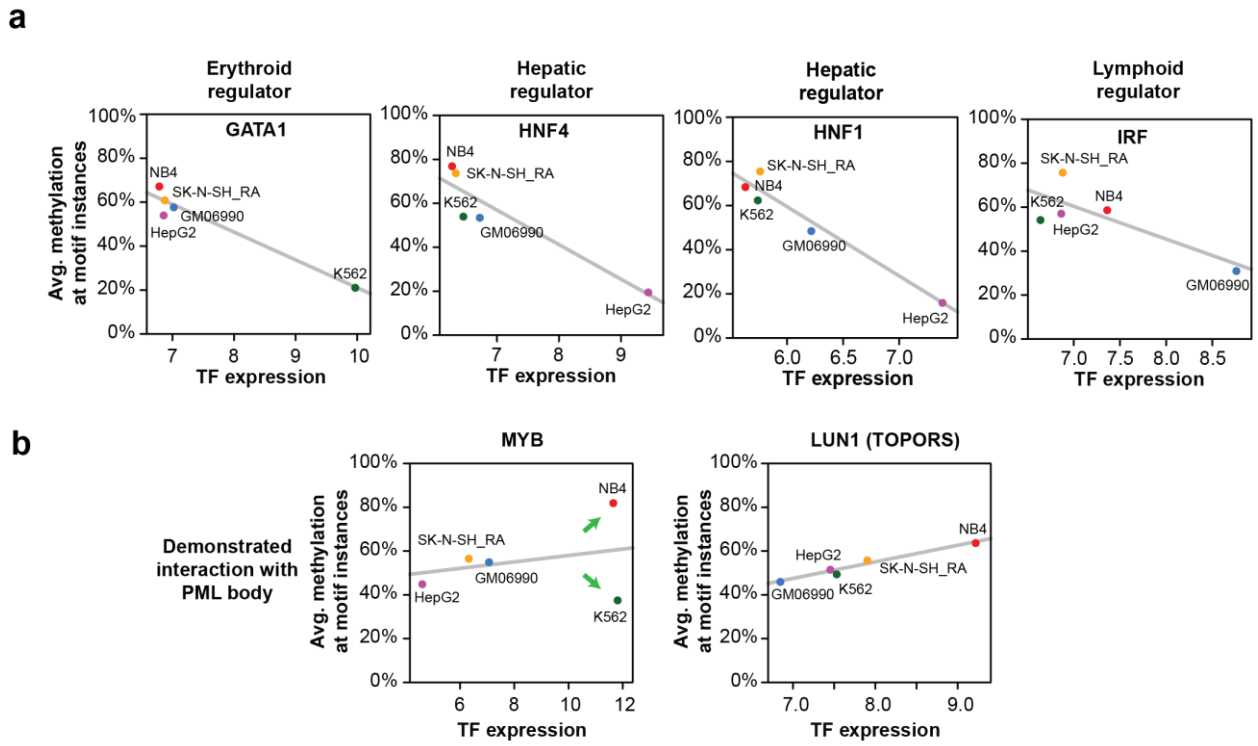
Supplementary Figure 9. This is a refinement of Fig. 3d. The top pie charts are identical in both figures. The bottom two pie charts here show the breakdown of novel promoter predictions with regard to their overlap separately with Gencode CAGE cluster TSS (left), and RIKEN CAGE cluster TSS (right), both of which datasets are described in the Supplementary Methods.



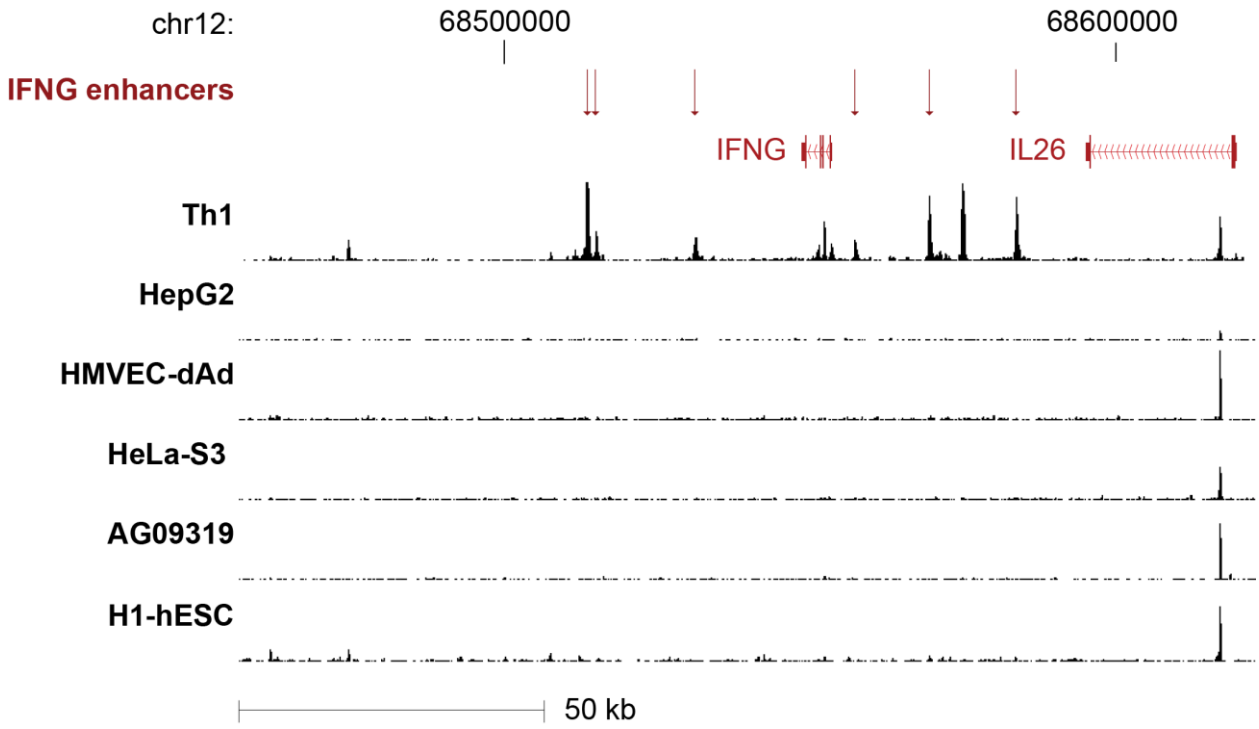
Supplementary Figure 10. Additional examples of novel promoters identified in K562 cells. (a) Novel prediction confirmed by CAGE and ESTs. (b) Novel prediction confirmed by CAGE annotation, no ESTs. (c), (d) Antisense promoter predictions at 3' end of annotated genes. (e) Antisense promoter prediction within Gencode-annotated genes.



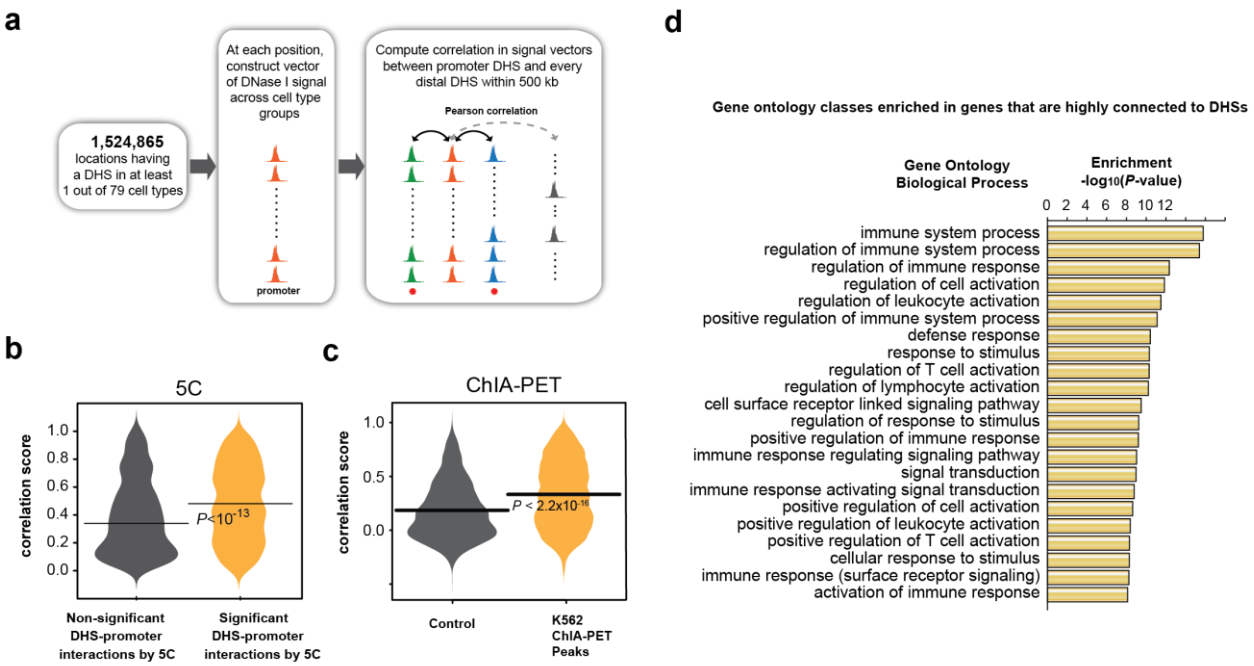
Supplementary Figure 11. (a) Further examples of association between methylation and accessibility. Data tracks show DNase I sensitivity in selected cell types. Green bars, CpG is 0% methylated; yellow, 50% methylated; red, 100% methylated. Association is quantified in the plots below the tracks. Each point in the graph represents one of 19 cell-types (a susbset of which is represented in the tracks). X-axis is the percent methylation of the site in that cell-type; y-axis is the normalized DNaseI tag density at the site in that cell type. In each example, accessibility (y-axis) quantitatively decreases as methylation increases (left to right). (b) Global characterization of the effect of methylation on chromatin accessibility, surveyed at 34,376 DHSs with RRBS data. 40% of sites with variable methylation across cell-types were associated with differences in chromatin accessibility. (c), In cell lines with methylated DHSs, site accessibility was reduced on average by 95%. Shown are sites where increased methylation was significantly associated with decreased accessibility (= 97% of all sites in the orange slice shown in (b)).



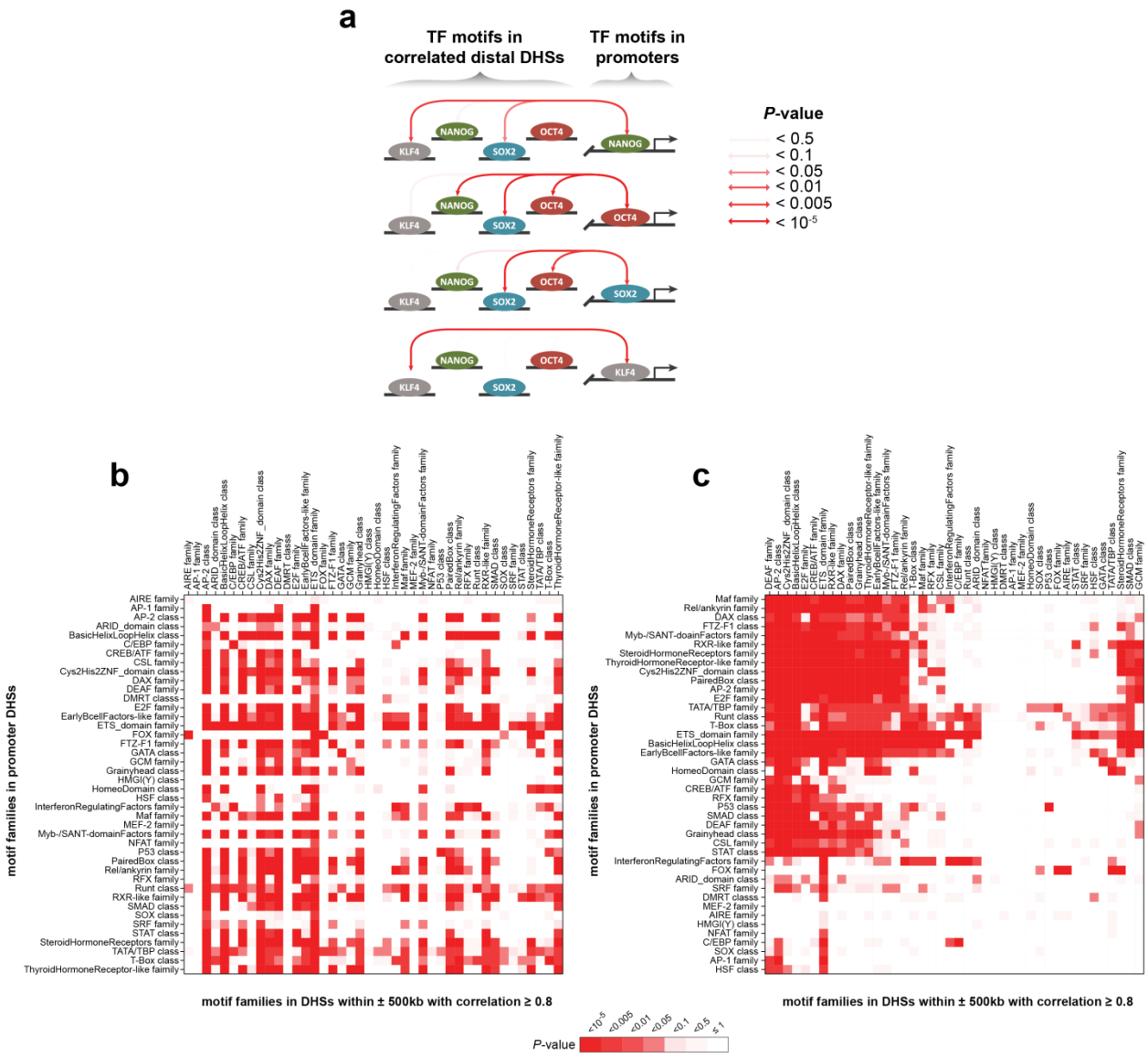
Supplementary Figure 12. (a) Relationship between TF transcript levels and overall methylation at cognate recognition sequences of the same TFs. Negative correlation indicates that site-specific DNA methylation follows TF vacation of differentially expressed TFs. Left, erythroid regulator in the erythroleukemia line K562; center, hepatic regulators in the liver carcinoma HepG2; and right, lymphoid regulator in the B lymphoblast line GM06990. (b), MYB and LUN1 have both been demonstrated to interact with PML bodies, and show increased transcription and binding site methylation in the acute promyelocytic leukemia (APL) line NB4. Although Myb expression is upregulated in both erythroid K562 and the APL line NB4 (green arrows), its putative binding sites exhibit altered methylation only in the APL line NB4.



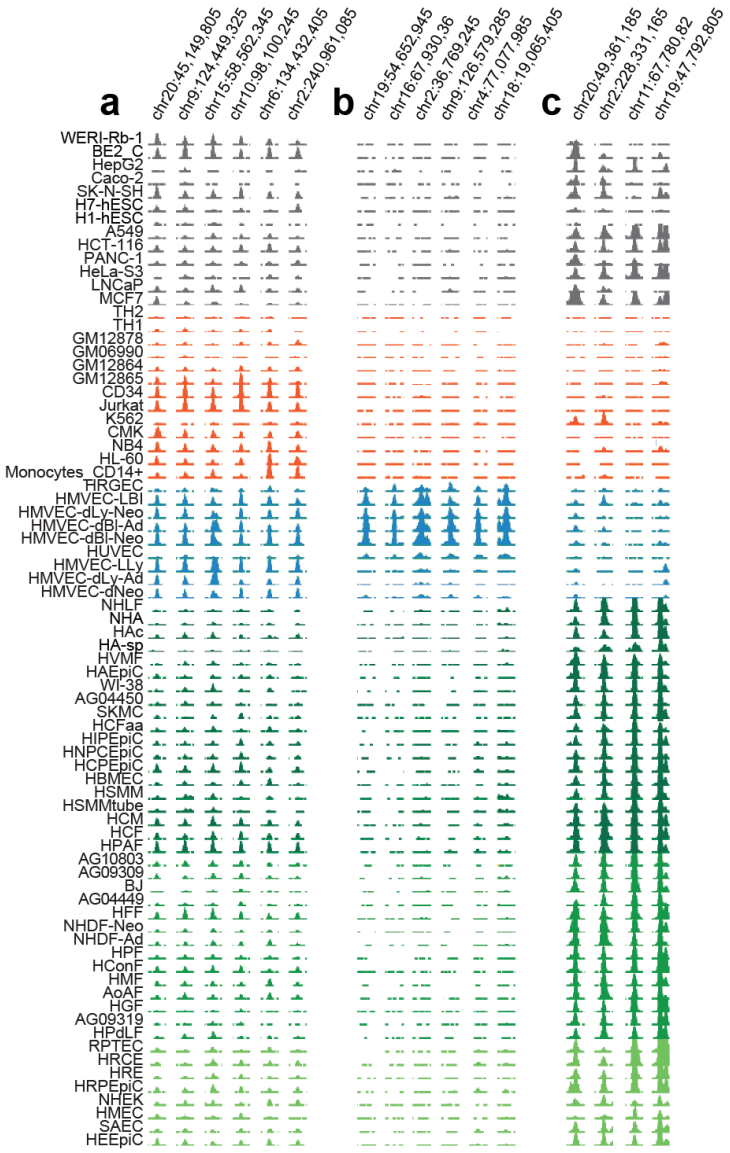
Supplementary Figure 13. Cell-specific enhancers (red arrows) in the IFNG locus. Enhancers of the IFNG gene[5] are marked by DHSs in the hTH1 (T lymphocyte) cell-type, consistent with the functioning of lymphocytes in producing the gene product interferon gamma. The enhancer loci are lacking in DHSs in other cell-types. Shown are DNaseI tag densities for six cell-types, including hTH1. See **Supplementary Table 4** for IFNG enhancer coordinates and references.



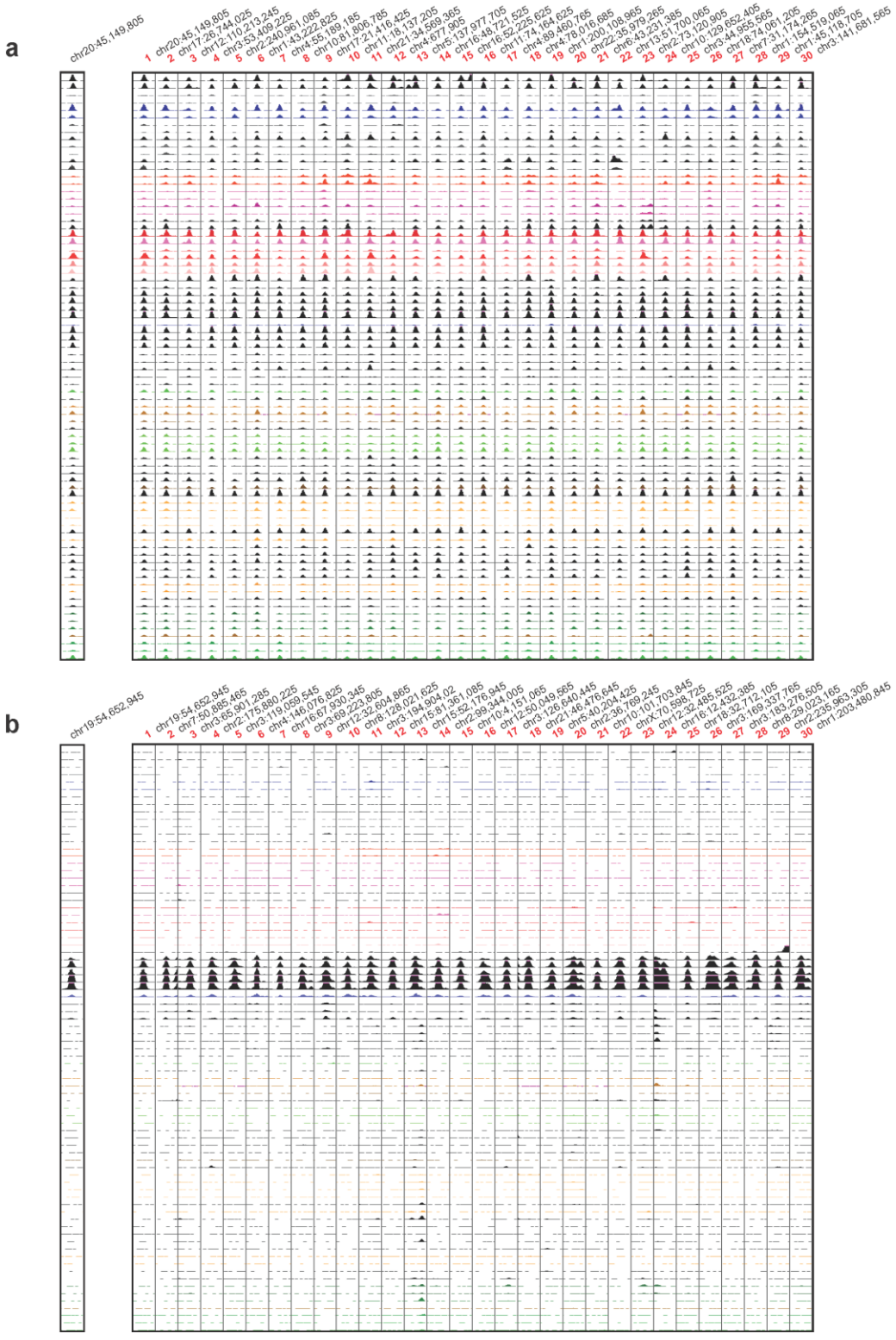
Supplementary Figure 14. Enrichments of 5C interactions, ChiaPET interactions, and gene ontology classes revealed by signal-vector correlation. (a) Each of 1,524,865 DHSs is treated as a vector of DNaseI densities across cell types. High correlations between vectors for promoter/distal DHS pairs separated by <500 kb identify DHSs likely co-regulated with specific promoters. (b) Distributions of maximal correlation scores for DHSs falling within independently ascertained peak interacting restriction fragments by 5C-seq (gold) vs. non-peak fragments (grey) for TSS-vs-all distal 5C-seq data collected over 1% of the human genome defined by ENCODE Pilot regions[6]. DHSs with high promoter correlation by cross-cell-type analysis show significantly increased chromatin interactions with the predicted cognate promoter ($P < 10^{-13}$). (c) Distribution of correlation scores for K562 ChiaPET[7] peak interactions in which both tags are in a K562 DHS and the tags are at least 10 kb apart (gold). Correlation scores for a random control set generated by scrambling the inter-tag distances while keeping the promoter tags fixed are shown in grey; as a group, these are significantly lower than the observed scores ($P < 2.2 \times 10^{-16}$). (d) Gene Ontology analysis performed on a list of all human genes with promoters connected to at least one DHS, ranked by the numbers of DHSs connected with each promoter. Shown is an unfiltered list of GO Biological Processes with $P < 10^{-8}$, indicating overwhelming enrichment of immune-related genes among genes with the most complex distal regulatory landscapes.



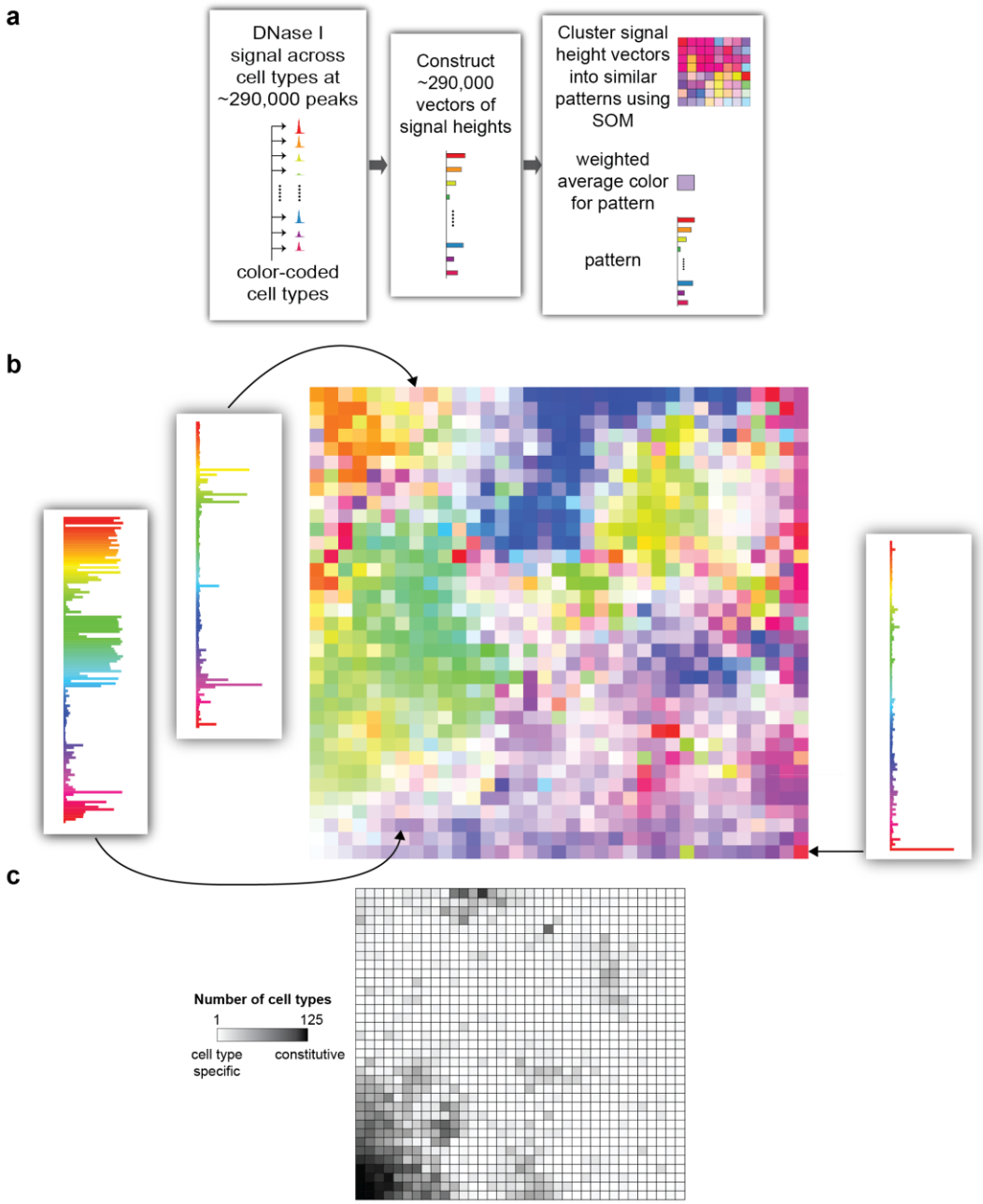
Supplementary Figure 15. Statistical significances of co-occurrences of motifs and families and classes of motifs within connected ($R > 0.8$) distal/promoter DHS pairs genome-wide. (a), Co-occurrences among motifs for pluripotency factors KLF4, SOX2, OCT4, and NANOG. Enriched co-occurrences are denoted by arrows shaded by *P*-value. (b)-(c), Co-occurrences of families and classes of motifs. Family and class definitions are given in **Supplementary Table 9**. In (b), the motif families and classes are shown in alphabetical order. The matrix is clearly not symmetric; for example, within co-occurrences, TATA/TBP is enriched in several cases when it appears in a promoter DHS, but in only a few cases when it appears in a correlated distal DHS. Panel (c) shows the data from (b), hierarchically clustered by column and row. The DAX, FTZ-F1, RXR-like, Steroid Hormone Receptors, and Thyroid Hormone Receptor-like families, which all belong to the same class, cluster tightly together by rows (presence within promoter DHSs).



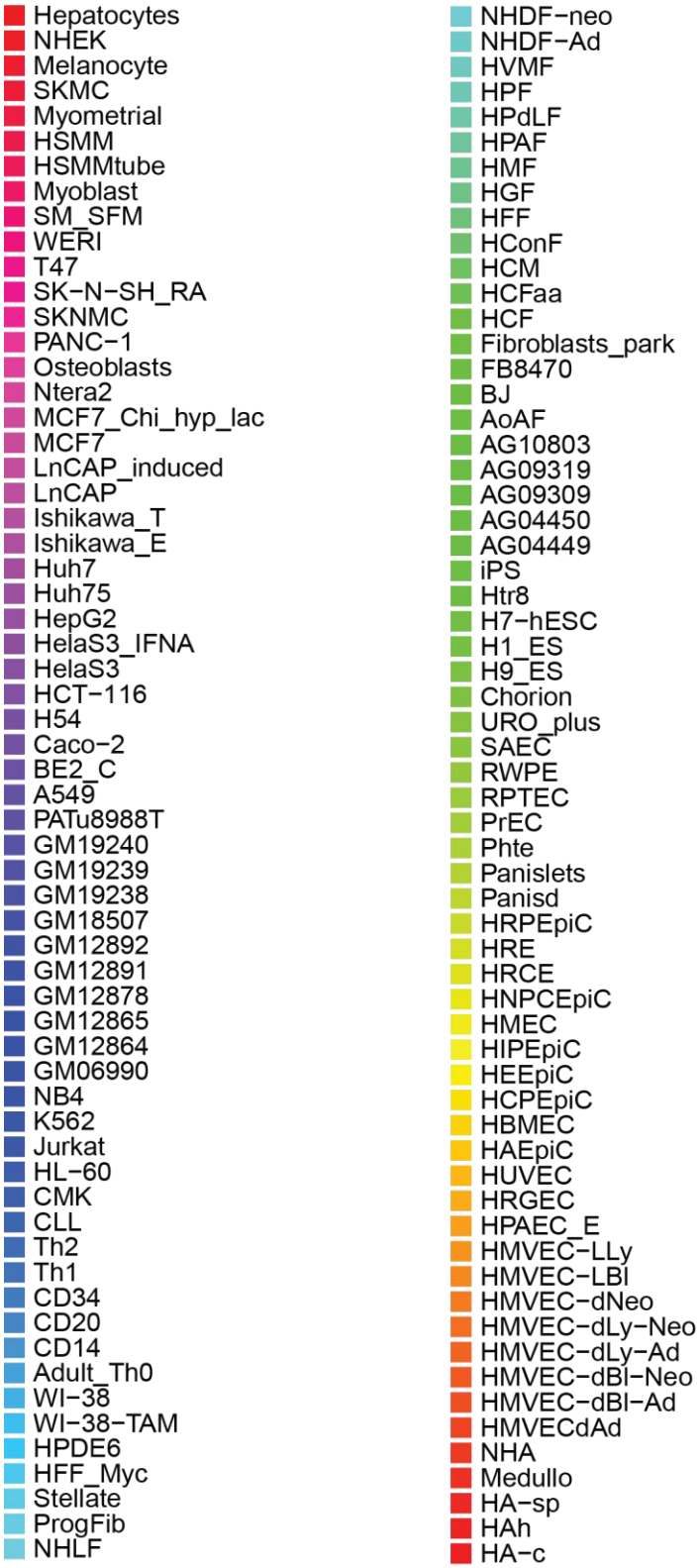
Supplementary Figure 16. (a)-(c), Examples of stereotyping of DHSs. In each case, a nearly identical cross-cell-type pattern of chromatin accessibility at DHS positions is observed for groups of DHSs widely separated *in trans*. Grey = immortal cells (pluripotent cells and cancer cell lines). Red = hematopoietic cells. Blue = endothelial cells. Green = epithelial, stromal cells, and visceral cells, with shading to denote different pattern groups.



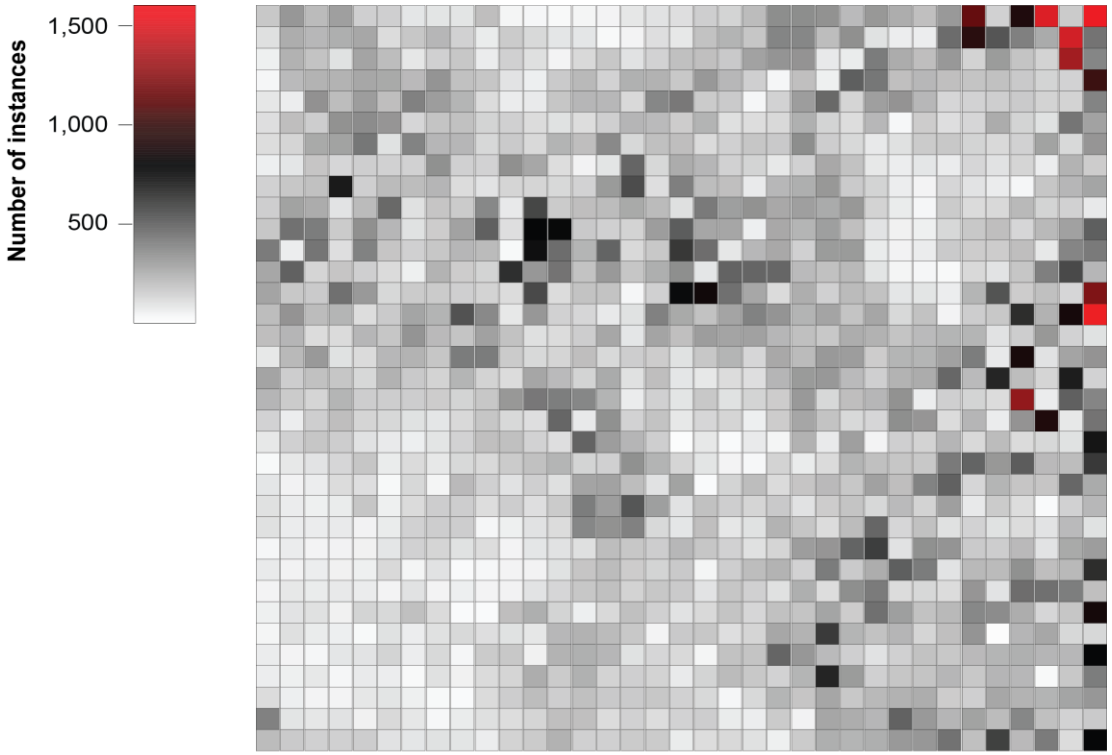
Supplementary Figure 17. (a) Top 30 ranked matches using our DNaseI pattern-matching algorithm (see Supplementary Methods) for the pattern in **Supplementary Fig. 16a**. (b) Top 30 matches for the pattern in **Supplementary Fig. 16b**. (Cell-type colouring in (a)-(b) does not match that in **Supplementary Fig. 16**.)



Supplementary Figure 18. Clustering of ~290,000 DHSs by cross-cell-type patterns using a self-organizing map (SOM), which learns patterns in the data and organizes DHSs into stereotyped groups analogous to those shown in Fig. 6a-e. (a) Schematic for SOM clustering and colour coding of patterns; index of cell types with their colours is given in **Supplementary Fig. 19**. (b) SOM of 1,225 DHS patterns. Each cell in the 35×35 grid represents one stereotyped pattern, with colour coding determined according to the weighted "average" cell type for that pattern. Three example pattern profiles are shown, corresponding to the indicated nodes in the grid. (c) Greyscale heatmap corresponding to that in (b) showing, for each colour-coded pattern, the cell-specificity of that pattern. Shading indicates cell-selectivity; black = DHS is constitutive (i.e. present in all cell types); white = DHS is cell type-specific; greyscale = gradations threreof. Note the concentration of patterns with promiscuous DHSs in the lower right; however, most stereotyped DHS patterns are highly cell-selective.



Supplementary Figure 19. Colour-coded key to the signal height vectors used as input for the SOM of **Supplementary Fig. 18**.



Code Number Map

172	328	213	301	149	151	85	86	88	204	44	41	27	42	46	80	93	106	152	136	210	360	360	344	215	327	261	202	320	1127	145	893	1494	164	1574
103	259	227	253	276	111	174	229	145	68	162	122	85	85	27	45	106	78	190	238	177	389	378	208	354	97	63	61	464	933	536	404	266	1458	441
57	246	188	96	284	206	80	148	176	66	146	103	48	120	174	93	142	196	201	169	292	339	343	74	68	416	263	203	324	138	312	184	137	1320	382
39	216	241	240	275	269	220	338	194	177	92	59	167	222	140	249	246	172	323	226	151	41	199	56	504	434	200	225	199	184	194	170	135	140	1007
82	61	347	207	307	204	403	307	236	111	66	73	177	229	227	120	380	428	163	167	36	140	307	474	130	302	342	227	128	141	152	166	138	138	392
107	204	158	346	365	337	128	185	195	134	146	111	96	110	151	231	215	161	238	103	109	120	250	135	110	227	33	159	105	127	263	134	96	435	292
108	132	270	288	434	97	394	229	229	124	165	119	237	204	374	124	198	205	115	147	178	168	360	158	123	76	100	164	171	135	120	145	288	121	335
59	85	183	127	153	144	145	350	147	141	354	275	109	49	86	486	125	264	228	145	135	78	144	274	192	79	63	83	65	127	95	132	62	240	161
145	175	194	715	149	212	206	232	111	91	130	131	122	142	316	566	105	407	207	200	73	225	255	325	165	133	63	46	127	91	64	41	179	229	283
152	288	264	90	214	475	100	182	162	376	74	584	162	205	209	271	213	102	424	303	345	274	301	266	325	138	70	40	164	117	118	221	133	143	75
174	448	414	163	353	229	97	159	291	503	107	788	777	170	147	138	307	510	278	280	58	272	234	304	200	91	52	55	126	173	204	129	224	294	504
419	62	436	107	399	183	147	176	229	213	30	753	461	235	491	118	172	627	457	80	225	273	146	314	317	69	50	66	166	161	161	203	77	386	423
276	499	131	182	150	115	55	241	161	133	655	330	443	187	319	179	254	359	85	492	487	482	217	193	224	86	42	38	29	120	89	180	407	574	228
286	172	167	452	321	126	163	164	150	138	111	573	114	187	189	41	146	768	845	452	277	117	256	147	213	191	83	88	62	234	543	161	196	128	1205
206	342	209	236	110	165	283	237	546	372	78	330	143	127	76	96	402	269	190	285	398	339	239	285	185	278	121	192	45	155	171	662	243	864	1603
204	231	111	115	229	180	338	219	184	294	191	104	113	158	277	84	284	200	307	285	294	230	197	171	259	245	184	216	235	235	93	156	326	140	98
77	224	319	97	116	227	240	178	418	417	162	122	132	204	122	158	160	97	175	234	127	272	212	319	306	159	234	236	294	414	75	868	94	284	342
286	157	193	177	125	243	170	137	236	138	114	280	216	136	265	100	204	85	78	111	107	233	318	323	216	150	242	262	479	215	685	208	132	708	140
231	163	132	236	155	157	155	161	136	176	325	430	411	380	237	65	148	87	172	98	134	159	319	125	203	229	346	61	149	134	112	1263	65	501	385
142	63	152	204	130	124	97	84	118	179	123	105	483	76	350	141	187	77	128	132	59	161	278	189	121	128	357	335	114	232	66	341	896	78	442
78	148	178	157	94	73	59	86	141	199	197	141	106	494	306	233	152	18	79	25	72	43	244	72	300	45	151	137	232	119	210	116	126	97	736
30	78	91	128	177	67	115	79	121	118	163	194	235	134	146	352	238	132	86	47	33	137	67	117	239	197	180	172	425	489	326	514	220	208	624
84	49	61	113	107	32	125	43	219	146	93	150	103	284	291	208	95	288	20	134	207	115	182	220	174	295	205	396	500	128	235	181	183	481	268
101	64	56	68	180	57	116	156	155	110	156	106	162	414	308	533	303	94	194	152	132	169	93	159	184	125	216	139	302	76	162	110	22	144	227
101	64	64	72	50	68	143	157	156	39	55	94	114	378	346	402	121	84	201	82	96	242	178	113	211	482	128	195	99	140	96	113	91	216	98
39	65	45	51	54	84	144	130	96	108	45	84	93	103	207	175	158	225	180	144	99	131	308	336	491	609	93	350	339	166	155	220	89	269	307
72	50	61	54	56	118	42	94	25	34	47	49	97	183	210	160	134	170	43	152	206	241	141	417	165	260	506	416	90	222	212	352	205	221	663
45	77	67	76	64	45	94	46	59	47	50	90	127	175	198	139	174	166	89	179	146	101	255	110	359	420	113	124	113	334	89	457	452	410	199
56	92	100	53	135	123	196	117	24	25	203	253	134	60	120	196	146	94	119	186	128	194	186	283	160	453	305	192	224	377	364	261	128	180	858
42	102	69	59	31	96	61	99	122	106	98	82	225	246	212	173	48	169	176	106	132	277	241	620	235	192	231	130	200	340	15	231	276	126	124
64	45	39	85	90	136	62	123	38	163	173	53	242	252	230	96	127	118	44	190	184	486	328	217	151	94	147	191	172	218	234	250	233	227	789
63	55	78	38	71	62	28	99	52	117	132	260	164	248	137	150	139	71	135	213	171	140	236	688	313	148	179	185	228	173	233	226	29	173	414
112	59	48	45	80	86	56	30	23	115	155	207	153	198	168	167	175	145	122	41	122	234	182	280	173	224	244	260	161	151	276	242	239	281	328
403	87	82	93	53	120	82	33	68	103	109	130	199	201	112	144	108	141	155	67	62	162	50	279	259	230	495	319	231	289	209	379	197	141	167
211	167	146	148	156	95	84	98	86	63	139	164	245	146	147	162	163	68	110	118	100	55	164	254	235	253	320	330	446	198	603	218	420	113	791

Supplementary Figure 20. The number of instances of each pattern discovered by the SOM illustrated in **Supplementary Fig. 18**; the top matrix is simply a heatmap version of the numeric matrix underneath.

Supplementary Table 1. Summary of all 125 cell-types for which DNaseI analysis was performed. Column 1 gives the abbreviated name as found in the figures, while column 2 gives a fully descriptive name. Column 3 indicates whether the DNase I data was collected by UW, Duke or both. Column 4 ("H" for "H3K4me3") indicates those cell-types for which H3K4me3 data was also available and used for promoter predictions or other analysis ("Y") or not ("N"). Column 5 ("S" for "sex") gives the sex of the donor(s): M, male, F, female, B, both sexes, U, undetermined.

Cell line	Description	Lab	H	S	Source	Cell/Tissue Protocol
A549	epithelial cell line derived from a lung carcinoma tissue	Duke/UW	Y	M	ATCC CCl-185	http://genome.ucsc.edu/ENCODE/protocols/cell/human/A549_Stam_protocol.pdf
GM12878	lymphoblastoid	Duke/UW	Y	F	Coriell GM12878	http://genome.ucsc.edu/ENCODE/protocols/cell/human/GM12878_protocol.pdf
HESC	H1 Human Embryonic Stem Cells	Duke/UW	N	M	Cellular Dynamics	http://genome.ucsc.edu/ENCODE/protocols/cell/human/H1_ES_protocol.pdf
HeLa-S3	cervical carcinoma	Duke/UW	Y	F	ATCC CCL-2.2	http://genome.ucsc.edu/ENCODE/protocols/cell/human/HeLa-S3_protocol.pdf
HepG2	liver carcinoma	Duke/UW	Y	M	ATCC HB-8065	http://genome.ucsc.edu/ENCODE/protocols/cell/human/HepG2_protocol.pdf
HMEC	Human Mammary Epithelial Cells	Duke/UW	Y	F	Lonza CC-3150	http://genome.ucsc.edu/ENCODE/protocols/cell/human/HMEC_Stam_protocol.pdf
HSMM	Normal Human Skeletal Muscle Myoblasts	Duke/UW	N	B	Lonza CC-2580	http://genome.ucsc.edu/ENCODE/protocols/cell/human/HSMM_Stam_protocol.pdf
HSMM tube	Normal Human Skeletal Muscle Myoblasts	Duke/UW	N	B	Lonza CC-2580	http://genome.ucsc.edu/ENCODE/protocols/cell/human/HSMM_Stam_protocol.pdf
HUVEC	Human Umbilical Vein Endothelial Cell	Duke/UW	Y	U	Lonza CC-2517	http://genome.ucsc.edu/ENCODE/protocols/cell/human/HUVEC_Stam_protocol.pdf
K562	leukemia	Duke/UW	Y	F	ATCC CCL-243	http://genome.ucsc.edu/ENCODE/protocols/cell/human/K562_protocol.pdf
LNCaP	prostate adeno-carcinoma	Duke/UW	Y	M	ATCC CRL-1740	http://genome.ucsc.edu/ENCODE/protocols/cell/human/LNCaP_Stam_protocol.pdf
MCF-7	mammary gland, adeno-carcinoma	Duke/UW	Y	F	ATCC HTB-22	http://genome.ucsc.edu/ENCODE/protocols/cell/human/Stam_15_protocols.pdf
Th1	primary human Th1 T cells	Duke/UW	N	U	primary pheresis of single normal subject	http://genome.ucsc.edu/ENCODE/protocols/cell/human/Stam_15_protocols.pdf
NHEK	Normal Human Epidermal Keratinocytes	Duke/UW	Y	F	Lonza CC-2501	http://genome.ucsc.edu/ENCODE/protocols/cell/human/Keratinocyte_protocol.pdf
AG04449	Fetal buttock/thigh fibroblast	UW	Y	M	Coriell AG04449	http://genome.ucsc.edu/ENCODE/protocols/cell/human/AGO4449_Stam_protocol.pdf
AG04450	Fetal lung fibroblast	UW	Y	M	Coriell AG04450	http://genome.ucsc.edu/ENCODE/protocols/cell/human/AG04450_Stam_protocol.pdf
AG09309	Adult human toe fibroblast	UW	Y	F	Coriell AG09309	http://genome.ucsc.edu/ENCODE/protocols/cell/human/AG09309_Stam_protocol.pdf
AG09319	Adult human gum tissue fibroblasts	UW	Y	F	Coriell AG09319	http://genome.ucsc.edu/ENCODE/protocols/cell/human/AG09309_Stam_protocol.pdf
AG10803	Adult human abdominal skin fibroblasts	UW	Y	M	Coriell AG10803	http://genome.ucsc.edu/ENCODE/protocols/cell/human/AG10803_Stam_protocol.pdf
AoAF	Normal Human Aortic Adventitial Fibroblast Cells	UW	Y	F	Lonza CC-7014, CC-7014T75	http://genome.ucsc.edu/ENCODE/protocols/cell/human/AoAF_Stam_protocol.pdf

Cell line	Description	Lab	H	S	Source	Cell/Tissue Protocol
BE2_C	Human neuroblastoma	UW	Y	M	ATCC CRL-2268	http://genome.ucsc.edu/ENCODE/protocols/cell/human/BE2-C_Stam_protocol.pdf
BJ	skin fibroblast	UW	Y	M	ATCC CRL-2522	http://genome.ucsc.edu/ENCODE/protocols/cell/human/BJ-tert_Stam_protocol.pdf
Caco-2	colorectal adeno-carcinoma	UW	Y	M	ATCC HTB-37	http://genome.ucsc.edu/ENCODE/protocols/cell/human/Stam_15_protocols.pdf
CMK	Human Acute Megakaryocytic Leukemia Cells	UW	N	M	DSMZ ACC-392	http://genome.ucsc.edu/ENCODE/protocols/cell/human/CMK_Stam_protocol.pdf
GM06990	B-Lymphocyte	UW	Y	F	Coriell GM06990	http://genome.ucsc.edu/ENCODE/protocols/cell/human/Stam_15_protocols.pdf
GM12864	B-Lymphocyte	UW	Y	M	Coriell GM12864	http://genome.ucsc.edu/ENCODE/protocols/cell/human/GM12864_Stam_protocol.pdf
GM12865	B-Lymphocyte	UW	Y	F	Coriell GM12865	http://genome.ucsc.edu/ENCODE/protocols/cell/human/GM12865_Stam_protocol.pdf
H7-hESC	Un-differentiated human embryonic stem cells	UW	Y	U	WiCell WA07(H7)	http://genome.ucsc.edu/ENCODE/protocols/cell/human/H7-hESC_Stam_protocol.pdf
HAc	Human Astrocytes-cerebellar	UW	Y	U	ScienCell 1810	http://genome.ucsc.edu/ENCODE/protocols/cell/human/HAc_Stam_protocol.pdf
HAEpiC	Human Amniotic Epithelial Cells	UW	N	U	ScienCell 7100	http://genome.ucsc.edu/ENCODE/protocols/cell/human/HAEpiC_Stam_protocol.pdf
HAh	Human Astrocytes - hippocampal	UW	N	F	ScienCell 1830	http://genome.ucsc.edu/ENCODE/protocols/cell/human/HAh_Stam_protocol.pdf
HA-sp	Human astrocytes spinal cord	UW	Y	U	ScienCell 1820	http://genome.ucsc.edu/ENCODE/protocols/cell/human/HA-sp_Stam_protocol.pdf
HBMEC	Human Brain Microvascular Endothelial Cells	UW	Y	U	ScienCell 1000	http://genome.ucsc.edu/ENCODE/protocols/cell/human/HBMEC_Myers_protocol.pdf
HCF	Human Cardiac Fibroblasts	UW	Y	U	ScienCell 6300	http://genome.ucsc.edu/ENCODE/protocols/cell/human/HCF_Stam_protocol.pdf
HCFaa	Human Cardiac Fibroblasts-Adult Atrial	UW	Y	F	ScienCell 6320	http://genome.ucsc.edu/ENCODE/protocols/cell/human/HCFaa_Stam_protocol.pdf
HCM	Human Cardiac Myocytes	UW	Y	U	ScienCell 6200	http://genome.ucsc.edu/ENCODE/protocols/cell/human/HCM_Stam_protocol.pdf
HConF	Human Conjunctival Fibroblast	UW	N	U	ScienCell 6570	http://genome.ucsc.edu/ENCODE/protocols/cell/human/HConF_Stam_protocol.pdf
HCPEpiC	Human Choroid Plexus Epithelial Cells	UW	Y	U	ScienCell 1310	http://genome.ucsc.edu/ENCODE/protocols/cell/human/HCPEpiC_Stam_protocol.pdf
HCT-116	colorectal carcinoma	UW	Y	M	ATCC CCL-247	http://genome.ucsc.edu/ENCODE/protocols/cell/human/HCT116_Stam_protocol.pdf
HEEpiC	Human Esophageal Epithelial Cells	UW	Y	U	ScienCell 2700	http://genome.ucsc.edu/ENCODE/protocols/cell/human/HEEpiC_Stam_protocol.pdf
HFF	Human Foreskin Fibroblast	UW	Y	M	Dr. Torok-Storb, Fred Hutchison Cancer Research Center	http://genome.ucsc.edu/ENCODE/protocols/cell/human/HFF_Stam_protocol.pdf
HFF_Myc	Human Foreskin Fibroblast	UW	Y	M	Dr. Torok-Storb, Fred Hutchison Cancer Research Center	http://genome.ucsc.edu/ENCODE/protocols/cell/human/HFF_Stam_protocol.pdf
HGF	Human Gingival Fibroblasts	UW	N	U	ScienCell 2620	http://genome.ucsc.edu/ENCODE/protocols/cell/human/HGF_Stam_protocol.pdf
HIPEpiC	Human Iris Pigment Epithelial Cells	UW	N	U	ScienCell 6560	http://genome.ucsc.edu/ENCODE/protocols/cell/human/HIPEpiC_Stam_protocol.pdf

Cell line	Description	Lab	H	S	Source	Cell/Tissue Protocol
HL-60	Human promyelocytic leukemia cells	UW	Y	F	ATCC CCL-240	http://genome.ucsc.edu/ENCODE/protocols/cell/human/HL-60_Stam_protocol.pdf
HMF	Human Mammary Fibroblast	UW	N	F	ScienCell 7630	http://genome.ucsc.edu/ENCODE/protocols/cell/human/HMF_Stam_protocol.pdf
HMVEC-dAd	Adult Human Dermal Microvascular Endothelial Cells	UW	N	U	Lonza CC-2543	http://genome.ucsc.edu/ENCODE/protocols/cell/human/HMVECdAd_Stam_protocol.pdf
HMVEC-dBl-Ad	Normal Adult Human Blood Microvascular Endothelial Cells, Dermal-Derived	UW	N	F	Lonza CC-2811, CC-2811T75	http://genome.ucsc.edu/ENCODE/protocols/cell/human/HMVEC-dBl-Ad_Stam_protocol.pdf
HMVEC-dBl-Neo	Normal Neonatal Human Blood Microvascular Endothelial Cells, Dermal-Derived	UW	N	M	Lonza CC-2813, CC-2813T75	http://genome.ucsc.edu/ENCODE/protocols/cell/human/HMVEC-dBl-Neo_Stam_protocol.pdf
HMVEC-dLy-Ad	Normal Adult Human Blood Microvascular Endothelial Cells, Dermal-Derived	UW	N	F	Lonza CC-2810, CC-2810T75	http://genome.ucsc.edu/ENCODE/protocols/cell/human/HMVEC-dLy-Ad_Stam_protocol.pdf
HMVEC-dLy-Neo	Normal Neonatal Human Lymphatic Microvascular Endothelial Cells, Dermal-Derived	UW	N	M	Lonza CC-2812, CC-2812T25	http://genome.ucsc.edu/ENCODE/protocols/cell/human/HMVEC-dLy-Neo_Stam_protocol.pdf
HMVEC-dNeo	Normal Neonatal Human Microvascular Endothelial Cells (single Donnor), Dermal-Derived	UW	N	M	Lonza CC-2505, CC-2505T225	http://genome.ucsc.edu/ENCODE/protocols/cell/human/HMVEC-dNeo_Stam_protocol.pdf
HMVEC-LBl	Normal Human Blood Microvascular Endothelial Cells, Lung-Derived	UW	N	F	Lonza CC-2815, CC-2815T75	http://genome.ucsc.edu/ENCODE/protocols/cell/human/HMVEC-Lbl_Stam_protocol.pdf
HMVEC-LLy	Normal Human Lymphatic Microvascular Endothelial Cells, Lung-Derived	UW	N	F	Lonza CC-2814, CC-2814T25	http://genome.ucsc.edu/ENCODE/protocols/cell/human/HMVEC-LLy_Stam_protocol.pdf
HNPC-EpiC	Human Non-Pigment Ciliary Epithelial Cells	UW	N	U	ScienCell 6580	http://genome.ucsc.edu/ENCODE/protocols/cell/human/HNPCEpiC_Stam_protocol.pdf
HPAEC	Human Pulmonary Artery Endothelial Cells	UW	N	U	Lonza CC-2530	http://genome.ucsc.edu/ENCODE/protocols/cell/human/HPAEC_Stam_protocol.pdf

Cell line	Description	Lab	H	S	Source	Cell/Tissue Protocol
HPAF	Human Pulmonary Artery Fibroblasts	UW	Y	U	ScienCell 3120	http://genome.ucsc.edu/ENCODE/protocols/cell/human/HPAF_Stam_protocol.pdf
HPdLF	Normal Human Periodontal Ligament Fibroblast Cells	UW	N	M	ScienCell 7409	http://genome.ucsc.edu/ENCODE/protocols/cell/human/HPdLF_Stam_protocol.pdf
HPF	Human Pulmonary Fibroblasts	UW	Y	U	ScienCell 3300	http://genome.ucsc.edu/ENCODE/protocols/cell/human/HPF_Stam_protocol.pdf
HRCEpiC	Human Renal Cortical Epithelial cells (normal)	UW	N	U	Lonza CC-2554	http://genome.ucsc.edu/ENCODE/protocols/cell/human/HRCEpiC_Stam_protocol.pdf
HRE	Human Renal Epithelial cells (normal)	UW	Y	U	Lonza CC-2556	http://genome.ucsc.edu/ENCODE/protocols/cell/human/HRE_Stam_protocol.pdf
HRGEC	Human Renal Glomerular Endothelial Cells	UW	N	U	ScienCell 4000	http://genome.ucsc.edu/ENCODE/protocols/cell/human/HRGEC_Stam_protocol.pdf
HRPEpiC	Human Retinal Pigment Epithelial Cells	UW	Y	U	ScienCell 6540	http://genome.ucsc.edu/ENCODE/protocols/cell/human/HRPEpiC_Stam_protocol.pdf
HVMF	Human Villous Mesenchymal Fibroblast Cells	UW	Y	U	ScienCell 7130	http://genome.ucsc.edu/ENCODE/protocols/cell/human/HVMF_Stam_protocol.pdf
Jurkat	T lympho-blastoid cell line derived from an acute T cell leukemia	UW	Y	M	ATCC TIB-152	http://genome.ucsc.edu/ENCODE/protocols/cell/human/Jurkat_Stam_protocol.pdf
Monocytes-CD14+	Monocytes-CD14+ are CD14-positive cells from human leukapheresis product	UW	Y	F	S. Heimfeld Laboratory, Fred Hutchison Cancer Research Center	http://genome.ucsc.edu/ENCODE/protocols/cell/human/MonoCD14_Stam_protocol.pdf
NB4	acute promyelocytic leukemia cell line	UW	Y	U	Refer to protocol documents for differing sources	http://genome.ucsc.edu/ENCODE/protocols/cell/human/NB4_Stam_protocol.pdf
NH-A	normal human astrocytes	UW	N	U	Lonza CC-2565	http://genome.ucsc.edu/ENCODE/protocols/cell/human/
NHDF-Ad	Adult Normal Human Dermal Fibroblasts	UW	N	F	Lonza CC-2511, CC-2511T225	http://genome.ucsc.edu/ENCODE/protocols/cell/human/NHDF-Ad_Stam_protocol.pdf
NHDF-neo	Neonatal Human Dermal Fibroblasts	UW	Y	U	Lonza CC-2509	http://genome.ucsc.edu/ENCODE/protocols/cell/human/NHDF-neo_Stam_protocol.pdf
NHLF	Normal Human Lung Fibroblasts	UW	Y	U	Lonza CC-2512	http://genome.ucsc.edu/ENCODE/protocols/cell/human/NHLF_Stam_protocol.pdf
NT2-D1	Human malignant pluripotent embryonal cancer cell line - Induced by RA to neuronal		N	M	ATCC CRL-1973	http://genome.ucsc.edu/ENCODE/protocols/cell/human/NT2-D1_protocol.pdf
PANC-1	pancreatic carcinoma	UW	Y	M	ATCC CRL-1469	http://genome.ucsc.edu/ENCODE/protocols/cell/human/PANC-1_Stam_protocol.pdf
PrEC	Human Prostate Epithelial Cell Line (PrEC/NHPRE)	UW	N	M	Lonza CC-2555	http://genome.ucsc.edu/ENCODE/protocols/cell/human/PrEC_Stam_protocol.pdf

Cell line	Description	Lab	H	S	Source	Cell/Tissue Protocol
RPTEC	Renal Proximal Tubule Epithelial Cells	UW	Y	U	Lonza CC-2553, CC-2553T225	http://genome.ucsc.edu/ENCODE/protocols/cell/human/RPTEC_Stam_protocol.pdf
SAEC	Small Airway Epithelial Cells	UW	Y	U	Lonza CC-2547	http://genome.ucsc.edu/ENCODE/protocols/cell/human/SAEC_Stam_protocol.pdf
SKMC	Human Skeletal Muscle Cells	UW	Y	U	Lonza CC-2561	http://genome.ucsc.edu/ENCODE/protocols/cell/human/SkMC_Stam_protocol.pdf
SK_N_MC	Neuro-epithelioma cell line derived from a metastatic supra-orbital human brain tumor	UW	Y	F	ATCC HBT-10	http://genome.ucsc.edu/ENCODE/protocols/cell/human/SK-N-MC_Stam_protocol.pdf
SK-N-SH_RA	neuroblastoma cell line differentiated w/ retinoic acid	UW	Y	F	ATCC HTB-11	http://genome.ucsc.edu/ENCODE/protocols/cell/human/Stam_15_protocols.pdf
Th2	Primary human Th2 T cells	UW	N	U	None (primary pheresis of single normal subject)	http://genome.ucsc.edu/ENCODE/protocols/cell/human/Th2_Stam_protocol.pdf
WERI-Rb-1	retinoblastoma	UW	Y	F	ATCC HTB-169	http://genome.ucsc.edu/ENCODE/protocols/cell/human/WERI-Rb-1_Stam_protocol.pdf
WI-38	Embryonic Lung Fibroblast Cells, hTERT immortalized, includes Raf1 construct	UW	Y	F	Dr. Carl Mann, SBIGeM	http://genome.ucsc.edu/ENCODE/protocols/cell/human/WI38_Stam_protocol.pdf
WI-38_TAM	Embryonic lung fibroblasts immortilized hTERT - Tamoxifen treated	UW	Y	F	Dr. Carl Mann, SBIGeM	http://genome.ucsc.edu/ENCODE/protocols/cell/human/WI38_Stam_protocol.pdf
CD20	Human B Cells	UW	Y	F	S. Heimfeld Laboratory, Fred Hutchison Cancer Research Center	http://genome.ucsc.edu/ENCODE/protocols/cell/human/CD20+_Stam_protocol.pdf
CD34	Mobilized primary CD34-positive cells from human leukapheresis product	UW	N	F	S. Heimfeld Laboratory, Fred Hutchison Cancer Research Center	http://www.roadmapepigenomics.org/files/protocols/experimental/dnaseI_sensitivity/HematopoieticCells_DNaseTreatment_v5_UW-NREMC.pdf
Th0	Unstimulated Th0 cells isolated from Adults' blood	Duke	N	M	Dr. Robin Haton at University of Alabama	submitted
HSMM _emb	embryonic myoblast	Duke	N	U	Duke/UNC/UT/EBI ENCODE group Muscle needle biopsies	http://genome.ucsc.edu/ENCODE/protocols/cell/human/HSMMe_Crawford_protocol.pdf
Ishikawa/ Estradiol_ 10nM_ 30m	endometrial adeno-carcinoma cells treated with 10 nM 17-bestradiol for 30 min	Duke	N	F	SIGMA-ALDRICH 99040201	http://genome.ucsc.edu/ENCODE/protocols/cell/human/Ishikawa_Crawford_protocol.pdf
Ishikawa/ 4OHTAM_ 100nM_ 30m	endometrial adeno-carcinoma treated with 100 nM 4-OH Tamoxifen for 30 min	Duke	N	F	SIGMA-ALDRICH 99040201	http://genome.ucsc.edu/ENCODE/protocols/cell/human/Ishikawa_Crawford_protocol.pdf
RWPE1	Prostate epithelial	Duke	N	M	ATCC CRL-11609	http://genome.ucsc.edu/ENCODE/protocols/cell/human/RWPE1_Crawford_protocol.pdf

Cell line	Description	Lab	H	S	Source	Cell/Tissue Protocol
8988T	human pancreas adeno-carcinoma (PA-TU-8988T), "established in 1985 from the liver metastasis of a primary pancreatic adeno-carcinoma from a 64-year-old woman" - DSMZ	Duke	N	F	DSMZ ACC 162	http://genome.ucsc.edu/ENCODE/protocols/cell/human/8988T_Crawford_protocol.pdf
AoSMC/ serum_ free_ media	aortic smooth muscle cells treated in serum-free media for 36 h	Duke	N	U	Lonza CC-2571	http://genome.ucsc.edu/ENCODE/protocols/cell/human/AoSMC_Crawford_protocol.pdf
Chorion	chorion cells (outermost of two fetal membranes), fetal membranes were collected from women who underwent planned cesarean delivery at term, before labor and without rupture of membranes.	Duke	N	U	Dr. Amy Murtha at Duke University (Durham, NC)	http://genome.ucsc.edu/ENCODE/protocols/cell/human/Chorion_and_decidua_Crawford_protocol.pdf
CLL	chronic lymphocytic leukemia cell, T-cell lymphocyte	Duke	N	F	Dr. Jennifer Brown, Department of Medicine, Harvard Medical School	http://genome.ucsc.edu/ENCODE/protocols/cell/human/CLL_Crawford_protocol.pdf
Fibrobl	Normal child fibroblast	Duke	N	F	Coriell AG08470	http://genome.ucsc.edu/ENCODE/protocols/cell/human/fibroblast_Crawford_protocol.pdf
FibroP	normal fibroblasts taken from individuals with Parkinson's disease, AG20443, AG08395 and AG08396 were pooled for this sample	Duke	N	U	Paul Tesar at Case Western University	http://genome.ucsc.edu/ENCODE/protocols/cell/human/FibroP_Crawford_protocol.pdf

Cell line	Description	Lab	H	S	Source	Cell/Tissue Protocol
Gliobla	glioblastoma, these cells (aka H54 and D54) come from a surgical resection from a patient with glioblastoma multiforme (WHO Grade IV). D54 is a commonly studied glioblastoma cell line[8] that has been thoroughly described[9]	Duke	N	U	Duke University Medical Center, requests for D54 cells should be directed to Darrell Bigner	http://genome.ucsc.edu/ENCODE/protocols/cell/human/D54_Crawford_protocol.pdf
GM12891	B-Lymphocyte, Lympho-blastoid, International HapMap Project, CEPH/Utah pedigree 1463, Treatment: Epstein-Barr Virus transformed	Duke	N	M	Coriell GM12891	http://genome.ucsc.edu/ENCODE/protocols/cell/human/GM12891_Crawford_protocol.pdf
GM12892	B-Lymphocyte, Lympho-blastoid, International HapMap Project, CEPH/Utah pedigree 1463, Treatment: Epstein-Barr Virus transformed	Duke	N	F	Coriell GM12892	http://genome.ucsc.edu/ENCODE/protocols/cell/human/GM12892_Crawford_protocol.pdf
GM18507	Lympho-blastoid, International HapMap Project, Yoruba in Ibadan, Nigera, Treatment: Epstein-Barr Virus transformed	Duke	N	M	Coriell GM18507	http://genome.ucsc.edu/ENCODE/protocols/cell/human/GM18507_protocol.pdf
GM19238	B-Lymphocyte, Lympho-blastoid, International HapMap Project, Yoruba in Ibadan, Nigera, Treatment: Epstein-Barr Virus transformed	Duke	N	F	Coriell GM19238	http://genome.ucsc.edu/ENCODE/protocols/cell/human/GM19238_Crawford_protocol.pdf

Cell line	Description	Lab	H	S	Source	Cell/Tissue Protocol
GM19239	B-Lymphocyte, Lympho-blastoid, International HapMap Project, Yoruba in Ibadan, Nigera, Treatment: Epstein-Barr Virus transformed	Duke	N	M	Coriell GM19239	http://genome.ucsc.edu/ENCODE/protocols/cell/human/GM19239_Crawford_protocol.pdf
GM19240	B-Lymphocyte, Lympho-blastoid, International HapMap Project, Yoruba in Ibadan, Nigera, Treatment: Epstein-Barr Virus transformed	Duke	N	F	Coriell GM19240	http://genome.ucsc.edu/ENCODE/protocols/cell/human/GM19240_Crawford_protocol.pdf
H9ES	human embryonic stem cell (hESC) H9	Duke	N	F	WiCell WA09	http://genome.ucsc.edu/ENCODE/protocols/cell/human/BG02ES_and_H9ES_Myers_protocols.pdf
HeLa-S3/IFNa4h	cervical carcinoma treated with IFN-alpha for 4h	Duke	N	F	ATCC CCL-2.2	http://genome.ucsc.edu/ENCODE/protocols/cell/human/HeLa-S3_IFN_Crawford_protocol.pdf
Hepato-cytes	Primary Human Hepatocytes, liver perfused by enzymes to generate single cell suspension	Duke	N	B	Zin-Bio	http://genome.ucsc.edu/ENCODE/protocols/cell/human/Hepatocytes_Crawford_protocol.pdf
HPDE6-E6E7	normal human pancreatic duct cells immortalized with E6E7 gene of HPV	Duke	N	F	Dr. Ming-Sound Tsao, Ontario Cancer Institute	http://genome.ucsc.edu/ENCODE/protocols/cell/human/HPDE6-E6E7_Crawford_protocol.pdf
HTR8svn	Trophoblast (HTR-8/SVneo) cell line. A thin layer of ectoderm that forms the wall of many mammalian blastulas and functions in the nutrition and implantation of the embryo.	Duke	N	F	Dr. Charles H. Graham, Department of Anatomy & Cell Biology, Queen's University at Kingston, Kingston, Ontario, Canada HTR8svhttp://genome.ucsc.edu/ENCODE/protocols/cell/human/Trophobl_Crawford_protocol.pdf	http://genome.ucsc.edu/ENCODE/protocols/cell/human/HTR8svn_Crawford_protocol.pdf
Huh-7.5	Hepatocellular carcinoma, hepatocytes selected for high levels of hepatitis C replication	Duke	N	M	Dr. Ravi Jhaveri at Duke University	http://genome.ucsc.edu/ENCODE/protocols/cell/human/Huh-7.5_Crawford_protocol.pdf
Huh-7	Hepatocellular carcinoma	Duke	N	M	Dr. Ravi Jhaveri at Duke University	http://genome.ucsc.edu/ENCODE/protocols/cell/human/Huh-7_Crawford_protocol.pdf

Cell line	Description	Lab	H	S	Source	Cell/Tissue Protocol
iPS	induced pluripotent stem cell derived from skin fibroblast	Duke	N	B	Dr. Josh Chenoweth, Laboratory of Molecular Biology, National Institutes of Health	http://genome.ucsc.edu/ENCODE/protocols/cell/human/iPS_Crawford_protocol.pdf
LNCaP/ androgen	prostate adeno-carcinoma treated with androgen, "LNCaP clone FGC was isolated in 1977 by J.S. Horoszewicz, et al., from a needle aspiration biopsy of the left supraclavicular lymph node of a 50-year-old caucasian male (blood type B+) with confirmed diagnosis of metastatic prostate carcinoma." – ATCC.	Duke	N	M	ATCC CRL-1740	http://genome.ucsc.edu/ENCODE/protocols/cell/human/LNCaP_Crawford_protocol.pdf
MCF-7/ Hypoxia_ LacAcid	MCF7 cells treated with hypoxia and lactose	Duke	N	F	ECACC 86012803	http://genome.ucsc.edu/ENCODE/protocols/cell/human/MCF-7_Crawford_protocol.pdf
Medullo	Medullo-blastoma (aka D721), surgical resection from a patient with medullo-blastoma as described by Darrell Bigner (1997)	Duke	N	F	Darrell Bigner, Duke University Medical Center	http://genome.ucsc.edu/ENCODE/protocols/cell/human/D721_Crawford_protocol.pdf
Melano	epidermal melanocytes	Duke	N	U	ScienCell 2200	http://genome.ucsc.edu/ENCODE/protocols/cell/human/Melano_Crawford_protocol.pdf
Myometr	Myometrial cells	Duke	N	F	Dr. Jennifer Condon at Magee Women's Research Institute (Pittsburg, PA)	http://genome.ucsc.edu/ENCODE/protocols/cell/human/Myometr_Crawford_protocol.pdf
Osteobl	normal human osteoblasts (NHOst)	Duke	N	U	Lonza CC-2538	http://genome.ucsc.edu/ENCODE/protocols/cell/human/Osteoblast_Crawford_protocol.pdf
PanIsletD	Dedifferentiated human pancreatic islets from one of the sources for PanIslets	Duke	N	B	National Disease Research Interchange (NDRI). PanIsletD	http://genome.ucsc.edu/ENCODE/protocols/cell/human/PanIsletD_Crawford_protocol.pdf
PanIslets	human pancreatic islets	Duke	N	B	See protocol document	http://genome.ucsc.edu/ENCODE/protocols/cell/human/PanIslets_Crawford_protocol.pdf
pHTE	Primary Human Tracheal Epithelial Cells	Duke	N	U	Dr. Cal Cotton at Case Western Reserve University	http://genome.ucsc.edu/ENCODE/protocols/cell/human/pHTE_Crawford_protocol.pdf

Cell line	Description	Lab	H	S	Source	Cell/Tissue Protocol
ProgFib	fibroblasts, Hutchinson-Gilford progeria syndrome (cell line HGPS, HGADFN167, progeria research foundation)	Duke	N	M	Progeria Research Foundation HGADFN167	http://genome.ucsc.edu/ENCODE/protocols/cell/human/progeria_Crawford_protocol.pdf
Stellate	Human Hepatic Stellate Cells, Liver that was perfused with collagenase and sellected for hepatic stellate cells by density gradient	Duke	N	U	Dr. Steve Choi at Duke University	http://genome.ucsc.edu/ENCODE/protocols/cell/human/Stellate_Crawford_protocol.pdf
T-47D	a human epithelial cell line derived from an mammary ductal carcinoma.	Duke	N	F	ATCC HTB-133	http://genome.ucsc.edu/ENCODE/protocols/cell/human/T47D_Myers_protocol.pdf
Urothelia	A primary culture of urothelial cells derived from a 12 year-old girl and immortalized by transfection with a temperature-sensitive SV-40 large T antigen gene, normal human ureter cells	Duke	N	F	lab of Dr. D Sens (University of N. Dakota) Urothelia	http://genome.ucsc.edu/ENCODE/protocols/cell/human/Urothelia_Crawford_protocol.pdf
Urothelia/UT189	Urotsa infected by UT189	Duke	N	F	lab of Dr. D Sens (University of N. Dakota) Urothelia	http://genome.ucsc.edu/ENCODE/protocols/cell/human/Urothelia_Crawford_protocol.pdf

Supplementary Table 2. Table showing overlap of repeat-masked elements by repeat family for families with more than 5000 elements overlapping DHSs. Column 1 shows the repeat family; column 2 shows the repeat class. Column 3 shows the average size of elements in the family; column 4 shows the total number of occurrences of elements of the family in the genome. Column 5 indicates the number of DHSs which overlap a member of the family by at least 50%, and Column 6 indicates the number of DHSs which overlap a member of the family by 100%.

Repeat family	Repeat class	Mean element size (bp)	# occurrences	# DHSs 50% overlapping	# DHSs 100% overlapping
hAT-Charlie	DNA	178.76	251950	47580	13234
hAT-Tip100	DNA	218.14	30241	6406	2704
CR1	LINE	178.42	60830	12992	4594
L1	LINE	544.91	938484	205129	145630
L2	LINE	225.16	462077	128696	61890
ERV1	LTR	482.95	172893	85365	63858
ERVK	LTR	845.34	10490	8025	7178
ERVL	LTR	356.05	157992	65237	41841
ERVL-MaLR	LTR	322.29	343675	110659	69172
Alu	SINE	260.93	1175329	71262	23399
MIR	SINE	142.79	590625	104043	15669
Low_complexity	Low_complexity	46.08	368110	6287	903
Simple_repeat	Simple_repeat	63.04	413687	9334	2116

Supplementary Table 3. List of DHS peaks with at least 50% overlap with Repeat-Masked sequence which were tested and found to be enhancers in transient assays (Supplementary Methods). begpos, starting coordinate of the element on the given chromosome; endpos, ending coordinate of the element.

Chromosome	DHS peak begpos	DHS peak endpos	Repetitive element begpos	Repetitive element endpos	Repetitive element name	Repetitive element family	Repetitive element class
chr1	22231480	22231630	22231559	22231642	L2a	L2	LINE
chr1	151569025	151569175	151568917	151569300	L2c	L2	LINE
chr1	151569180	151569330	151568917	151569300	L2c	L2	LINE
chr2	169708420	169708570	169708231	169708745	MLT1F2	ERVL-MaLR	LTR
chr5	56300505	56300655	56300471	56300691	L2c	L2	LINE
chr6	41691040	41691190	41691079	41691182	(CA)n	Simple_repeat	Simple_repeat
chr7	20259520	20259810	20259517	20259978	MLT1N2	ERVL-MaLR	LTR
chr7	116418000	116418150	116417992	116418227	Tigger15a	TcMar-Tigger	DNA
chr7	116418160	116418310	116417992	116418227	Tigger15a	TcMar-Tigger	DNA
chr8	144973800	144973950	144973885	144974179	MLT1I	ERVL-MaLR	LTR
chr9	131901965	131902115	131902049	131902190	MIR3	MIR	SINE
chr9	90925320	90925470	90925333	90925647	FordPrefect	hAT-Tip100	DNA
chr13	108594500	108594650	108593029	108598435	L1PA15-16	L1	LINE
chr14	24082720	24082870	24082518	24082816	AluJr4	Alu	SINE
chr14	24163800	24163950	24162344	24164444	HERV3-int	ERV1	LTR
chr15	96817040	96817190	96817070	96817269	L2b	L2	LINE
chr21	30850360	30850510	30850296	30850848	MLT2A2	ERVL	LTR
chr21	34752845	34752995	34752726	34752909	MER34C2	ERV1	LTR
chr21	34753360	34753510	34753330	34753651	L1MB7	L1	LINE
chr21	34753780	34753930	34753663	34753983	AluJb	Alu	SINE
chr21	35028340	35028490	35028404	35028630	MLT1K	ERVL-MaLR	LTR

Supplementary Table 4. A list of 1046 known regulatory elements, enhancers, LCRs, insulators, and silencers, with references. Due to the size of this file, we are making it available through the EBI ftp server at
ftp://ftp-private.ebi.ac.uk/byDataType/openchrom/jan2011/enhancers/literature_regulatory_elements.xls

This Excel file contains 1046 rows of data. Columns A-C contain the genomic coordinates (hg19); column D contains either the regulated gene, nearest gene, or an element name; and column E contains references in the literature for the element. The first five lines of data are shown below.

chr1	3190581	3191428	element_705	http://enhancer.lbl.gov
chr1	8130439	8131887	element_1833	http://enhancer.lbl.gov
chr1	10732070	10733118	element_289	http://enhancer.lbl.gov
chr1	10781239	10781744	element_389	http://enhancer.lbl.gov
chr1	10795106	10799241	element_2094	http://enhancer.lbl.gov

Supplementary Table 5. Manually-curated mapping between TRANSFAC motif models and gene names. Due to the size of this file, we are making it available through the EBI ftp server at ftp://ftp-private.ebi.ac.uk/byDataType/openchrom/jan2011/motif_to_gene/ Supplemental_Table_Fig6-Methylation_xfac2geneName.xls.

This Excel file contains 944 lines of data. The first five lines of data are shown below.

XFAC_MOTIF	GENE_SYMBOL
AHRARNT_01	AHR
AHRARNT_01	ARNT
AHRARNT_02	AHR
AHRARNT_02	ARNT
AHRHIF_Q6	AHR

Supplementary Table 6. Grouping of 79 cell types into 32 cell-type categories, for exploration of *cis*-connectivity among DHSs. The grouping was obtained by hierarchically clustering the cell types by their DHS locations across the genome. Descriptions of the cell types are given in **Supplementary Table 1**.

Category number	Cell types assigned to category
1	WERI_Rb1
2	BE_2_C
3	CACO2, HEPG2, SKNSH
4	HESC, hESCT0
5	A549, HCT116, Hela, PANC1
6	LNCap, MCF7
7	CD56, CD4, hTH1, hTH2
8	GM06990, GM12864, GM12865, GM12878
9	CD34, Jurkat
10	K562, CMK
11	NB4, HL60, CD14
12	HRGEC, HMVEC_LBl, HMVEC_dLyNeo, HMVEC_dBlAd, HMVEC_dBlNeo, HUVEC
13	HMVEC_LLy, HMVEC_dLyAd, HMVEC_dNeo
14	NHLF, NHA
15	HAc
16	HAsp
17	HVMF
18	HAEpiC
19	WI_38, AG04450, IMR90
20	SkMC
21	HCFaa
22	HIPEpiC, HNPCEpiC, HCPEpiC, HBMEC
23	HSMM, HSMM_D
24	HCM, HCF, HPAF
25	AG10803, AG09309, BJ, AG04449, HFF
26	NHDF_Neo, NHDF_Ad
27	HPF, HConF, HMF, AoAF
28	HGF, AG09319, HPdLF
29	RPTEC, HRCE, HRE
30	HRPEpiC
31	HMEC, NHEK
32	SAEC, HEEpiC

Supplementary Table 7. Genomic coordinates of all promoter DHSs and distal, non-promoter DHSs within ±500kb correlated with them at threshold 0.7. Due to the size of this file, we are making it available through the EBI ftp server at
ftp://ftp-private.ebi.ac.uk/byDataType/openchrom/jan2011/dhs_gene_connectivity/
genomewideCorrs_above0.7_promoterPlusMinus500kb_withGeneNames_32celltypeCategories.bed8.gz

This compressed, tab-delimited text file contains 1,672,958 lines of data, for 63,318 distinct promoter DHSs that each have at least one distal DHS connected to it. Each promoter DHS overlaps a TSS, or is the nearest DHS to the TSS in the 5' direction; columns 1-3 contain each promoter DHS's genomic coordinates (hg19). The Gencode gene names are given in column 4. Because distinct gene names can be given to the same TSS, and because distinct TSSs can have the same nearby DHS called as its promoter DHS, data for each promoter DHS is repeated in this file roughly three times on average, with a different gene name for each repetition (there are 207,878 distinct combinations of promoter DHS + gene name in this file). Columns 5-7 contain the genomic coordinates for each distal, non-promoter DHS within 500kb of the promoter DHS given in columns 1-3 that achieves correlation ≥0.7 with it; the correlation between the promoter/distal DHS pair is given in column 8. Distal DHSs appear multiple times in the file when they achieve correlation ≥0.7 with multiple promoter DHSs. Using program sort-bed from the BEDOPS genomic data analysis software suite, from the command line within a Unix system, the set of 578,905 distal DHSs connected with at least one promoter DHS can be extracted into a file named "outfile" by executing the command

```
cut -f5-7 infile | sort-bed - | uniq > outfile
```

where "infile" represents the file
genomewideCorrs_above0.7_promoterPlusMinus500kb_withGeneNames_32celltypeCategories.bed8.

The first five lines of data are shown below.

chr1	66660	66810	AL627309.1	chr1	87640	87790	0.87171
chr1	66660	66810	AL627309.1	chr1	118840	118990	0.908176
chr1	66660	66810	AL627309.1	chr1	136960	137110	0.915177
chr1	66660	66810	AL627309.1	chr1	566760	566910	0.731457
chr1	96520	96670	RP11-34P13.8	chr1	237020	237170	0.786171

Supplementary Table 8. Gene sets and search terms used to identify Gene Ontology Biological Processes enriched within genes highly connected to distal DHSs.

Gene sets	Search terms
neural	"neur", "brain", "action potential", "astrocyte", "axon", "hippocampus", "spinal", "nervous", "dendrocyte", "cerebr", "perception", "nerve", "glial"
cardiovascular	"heart", "cardio", "angio", "artery", "cardiac", "circulat", "vascu", "vein", "venous", "blood pressure", "blood vessel"
kidney	"kidney", "neph", "urogen", "renal", "urete"
liver	"hepatic", "liver", "bile", "biliary"
lung	"lung", "pulmon", "bronch", "trachea", "alveol"
gut	"gut", "intesti", "stomach", "bowel", "jejeunum", "caecum", "digestive"
bone	"osteo", "BMP", "bone", "skelet", "chondrocyte", "ossification", "cartilage," "ossify"
lipid/adipose tissue	"lipid", "sterol", "glyceride", "phosphatidyl", "sphingo", "acylglycerol", "icosanoid", "steroid", "adipose", "fat"
muscle	"muscle", "muscular", "myosin"
hematological	“blood", "hemo", "myeloid"
dermal	“dermal”, “skin”
immune	"immune", "interleukin", "B cell", "T cell", "cytokine", "NF-kappa", "leukocyte", "lymphocyte", "interferon"

Supplementary Table 9. Groupings of TRANSFAC motifs into families and classes according to the structures of their associated proteins. "Classes" are composed of "families." Data adapted from http://www.edgar-wingender.de/huTF_classification.html.

Family or class	Motifs
AIRE family	AIRE_01, AIRE_02
AP-1 family	AP1FJ_Q2, AP1_01, AP1_C, AP1_Q2, AP1_Q2_01, AP1_Q4, AP1_Q6, AP1_Q6_01, ATF3_Q6, ATF4_Q2, ATF5_01, ATF_B, BACH1_01, BACH2_01, NFE2_01, XBP1_01, XBP1_02
AP-2 class	AP2ALPHA_01, AP2ALPHA_02, AP2ALPHA_03, AP2GAMMA_01, AP2_Q3, AP2_Q6, AP2_Q6_01
ARID Domain class	BDP1_01, MRF2_01
Basic Helix-Loop-Helix (bHLH) class	AHRARNT_01, AHRARNT_02, AHRHIF_Q6, AHR_01, AHR_Q5, AP4_01, AP4_Q5, AP4_Q6, AP4_Q6_01, ARNT_01, ARNT_02, CMYC_01, CMYC_02, DEC2_Q2, DEC_Q1, E12_Q6, E2A_Q2, E2A_Q6, E47_01, E47_02, EBOX_Q6_01, HAND1E47_01, HEB_Q6, HEN1_01, HEN1_02, HES1_Q2, HIF1_Q3, HIF1_Q5, HIF2A_01, HTF_01, MATH1_Q2, MAX_01, MAX_Q6, MYCMAX_01, MYCMAX_02, MYCMAX_03, MYCMAX_B, MYOD_01, MYOD_Q6, MYOD_Q6_01, MYOGNF1_01, NEUROD_02, NMYC_01, SREBP1_01, SREBP1_02, SREBP1_Q5, SREBP2_Q6, SREBP_Q3, SREBP_Q6, TAL1ALPHAE47_01, TAL1BETAE47_01, TAL1BETAITF2_01, TAL1_01, TAL1_Q6, TCF11MAFG_01, TCF11_01, TCF3_01, TCF4_01, USF_01, USF_02, USF_C, USF_Q6, USF_Q6_01
C/EBP	CEBPA_01, CEBPB_01, CEBPB_02, CEBPDELTA_Q6, CEBPGAMMA_Q6, CEBP_01, CEBP_C, CEBP_Q2, CEBP_Q2_01, CEBP_Q3, HLF_01, TEF1_Q6_03, TEF_01, TEF_Q6
CREB/ATF family	ATF1_Q6, ATF_01, CREBATF_Q6, CREBP1_Q2, CREB_02, CREB_Q2, CREB_Q2_01, CREB_Q4, CREB_Q4_01, CREM_Q6, TAXCREB_01, TAXCREB_02
CSL family	RBPJK_01
Cys2His2ZNF domain class	BCL6_01, BCL6_02, BCL6_Q3, BLIMP1_Q6, CIZ_01, CKROX_Q2, CTCF_01, CTCF_02, E4F1_Q6, EGR1_01, EGR2_01, EGR3_01, EGR_Q6, EVI1_01, EVI1_02, EVI1_03, EVI1_04, EVI1_05, EVI1_06, FKLF_Q5, FPM315_01, GFI1B_01, GFI1_01, GFI1_Q6, GKLF_02, GLI1_01, GLI1_Q2, GLI2_01, GLI3_01, GLI3_02, GLI3_Q5_01, GLI_Q2, GTF2IRD1_01, GZF1_01, HELIOSA_01, HELIOSA_02, HIC1_02, HIC1_03, IK1_01, IK2_01, IK3_01, IK_Q5, KAISO_01, KLF15_Q2, KROX_Q6, LYF1_01, MAZR_01, MAZ_Q6, MTF1_01, MTF1_Q4, MZF1_02, NRSF_01, NRSF_Q4, PLZF_02, REST_01, REX1_03, RREB1_01, SP1SP3_Q4, SP1_01, SP1_02, SP1_Q2_01, SP1_Q4_01, SP1_Q6, SP1_Q6_01, SP2_01, SP3_Q3, SP4_Q5, STAF_01, STAF_02, SZF11_01, TFIIA_Q6, TFIII_Q6, WT1_Q6, YY1_01, YY1_02, YY1_Q6, YY1_Q6_02, ZBP89_Q4, ZBRK1_01, ZF5_B, ZFX_01, ZIC1_01, ZIC2_01, ZIC3_01, ZID_01, ZNF219_01, ZNF515_01
DAX family	DAX1_01
DEAF family	DEAF1_01, DEAF1_02
DMRT class	DMRT1_01, DMRT2_01, DMRT3_01, DMRT4_01, DMRT7_01
E2F family	E2F1_Q3_01, E2F1_Q4_01, E2F1_Q6_01, E2F_01, E2F_03, E2F_Q3_01, E2F_Q4_01, E2F_Q6_01
Early B cell Factors-like family	EBF_Q6
ETS Domain family	CETS1P54_01, CETS1P54_02, CETS1P54_03, EHF_01, ELF1_Q6, ELF5_01, ELK1_01, ELK1_02, ELK1_03, ELK1_04, ERG_01, ESE1_Q3, ETS1_B, ETS2_B, ETS_Q4, FLI1_Q6, GABP_B, NERF_Q2, PU1_01, PU1_Q4, SAP1A_01, SPIB_01, TEL2_Q6
FOX family	FOXD3_01, FOXJ2_01, FOXJ2_02, FOXM1_01, FOXO1_01, FOXO1_02, FOXO1_Q5, FOXO3A_Q1, FOXO3_01, FOXO4_01, FOXO4_02, FOXP1_01, FOXP3_Q4, FOX_Q2, FREAC2_01, FREAC3_01, FREAC4_01, FREAC7_01, HFH3_01, HFH4_01, HFH8_01, HNF3ALPHA_Q6, HNF3A_01, HNF3B_01, HNF3_Q6, HNF3_Q6_01, WHN_B
FTZ-F1 family	LRH1_Q5, SF1_Q6_01
GATA class	GATA1_01, GATA1_02, GATA1_03, GATA1_04, GATA1_05, GATA1_06, GATA2_01, GATA2_02, GATA2_03, GATA3_01, GATA3_02, GATA3_03, GATA4_Q3, GATA6_01, GATA_C
GCM class	GCM_Q2
Grainyhead class	ALPHACP1_01, CP2_01, CP2_02, LBP9_01, MECP2_01, MECP2_02
HMGI(Y) class	HMGA2_01, HMGIY_Q3

Family or class	Motifs
HomeoDomain class	AFP1_Q6, ALX3_01, ALX4_01, ALX4_02, ARP1_01, ARX_01, BARHL1_01, BARHL2_01, BARX1_01, BARX2_01, BRN2_01, BRN3C_01, BRN4_01, CART1_01, CART1_02, CART1_03, CDPCR1_01, CDPCR3HD_01, CDPCR3_01, CDP_01, CDP_02, CDP_03, CDP_04, CDX1_01, CDX2_01, CDX2_Q5, CDX2_Q5_01, CDX_Q5, CRX_02, CRX_Q4, DLX1_01, DLX2_01, DLX3_01, DLX5_01, DLX7_01, EMX2_01, EN1_02, EN2_01, ESX1_01, EVX1_01, GBX2_01, GSH2_01, HB9_01, HMBOX1_01, HMX1_02, HMX3_02, HNF1B_01, HNF1_01, HNF1_02, HNF1_C, HNF1_Q6, HNF1_Q6_01, HNF6_Q6, HOMEZ_01, HOX13_01, HOX13_02, HOXA10_01, HOXA11_01, HOXA13_02, HOXA13_03, HOXA1_01, HOXA2_01, HOXA3_02, HOXA4_01, HOXA6_01, HOXA9_01, HOXB13_01, HOXB3_01, HOXB4_01, HOXB5_01, HOXB6_01, HOXB8_01, HOXB9_01, HOXC10_01, HOXC11_01, HOXC12_01, HOXC13_01, HOXC4_01, HOXC5_01, HOXC8_01, HOXC9_01, HOXD12_01, HOXD13_01, HOXD1_01, HOXD3_01, HOXD9_Q2, IPF1_01, IPF1_02, IPF1_03, IPF1_04, IPF1_05, IPF1_06, IPF1_Q4, IPF1_Q4_01, IRX2_01, IRX4_01, IRX5_01, IRXB3_01, ISL1_Q6, ISX_01, LHX3_01, LHX3_02, LHX4_01, LHX5_01, LHX61_01, LHX61_02, LHX8_01, LMX1B_01, LMX1_01, MEIS1AHOXA9_01, MEIS1BHOXA9_02, MEIS1_01, MEIS1_02, MEIS2_01, MOX1_01, MSX1_01, MSX1_02, MSX2_01, NANOG_01, NANOG_02, NCX_01, NCX_02, NKX21_01, NKX22_01, NKX22_02, NKX25_03, NKX25_Q5, NKX32_02, NKX3A_01, NKX3A_02, NKX61_01, NKX61_02, NKX61_03, NKX62_Q2, OCT1_01, OCT1_02, OCT1_03, OCT1_04, OCT1_05, OCT1_06, OCT1_07, OCT1_08, OCT1_B, OCT1_Q5_01, OCT1_Q6, OCT2_01, OCT2_02, OCT4_01, OCT4_02, OCT_C, OCT_Q6, OTP_01, OTX1_01, OTX2_01, OTX2_Q3, OTX3_01, PBX1_01, PBX1_02, PBX1_03, PBX1_04, PBX_Q3, PIT1_01, PIT1_Q6, PITX1_01, PITX1_Q6, PITX2_01, PITX2_Q2, PITX3_01, PITX3_Q2, PKNOX2_01, PMX2A_01, PMX2B_01, POU1F1_Q6, POU2F3_01, POU3F2_01, POU3F2_02, POU5F1_01, POU6F1_01, POU6F1_02, POU6F1_03, PREP1_01, PROP1_02, RAX_01, SATB1_Q3, SHOX2_01, SIX1_01, SIX2_01, SIX3_01, SIX4_01, SIX6_01, SIX6_02, TGIF2_01, TGIF_01, TGIF_02, TTF1_Q6, VAX1_01, VAX2_01, VSX1_01
HSF class	HSF1_01, HSF1_Q6, HSF2_01, HSF2_02, HSF_Q6
Interferon Regulating Factors family	ICSBP_Q6, IRF1_01, IRF2_01, IRF3_Q3, IRF7_01, IRF_Q6, IRF_Q6_01
Maf family	CMAF_01, LMAF_Q2, MAF_Q6, MAF_Q6_01
MEF-2 family	AMEF2_Q6, HMEF2_Q6, MEF2_01, MEF2_02, MEF2_03, MEF2_04, MEF2_05, MEF2_Q6_01, MMEF2_Q6, RSRFC4_01, RSRFC4_Q2
Myb-/SANT-domain Factors family	CDC5_01, CMYB_01, CMYB_Q5, MYB_Q3, MYB_Q5_01, MYB_Q6
NFAT family	NFAT2_01, NFAT3_Q3, NFAT_Q4_01, NFAT_Q6
P53 class	P53_01, P53_02, P53_03, P53_04, P53_05, P53_DECAMER_Q2, P63_01
PairedBox class	PAX1_B, PAX2_01, PAX2_02, PAX3_01, PAX3_B, PAX4_01, PAX4_02, PAX4_03, PAX4_04, PAX4_05, PAX5_01, PAX5_02, PAX6_01, PAX6_02, PAX6_Q2, PAX7_01, PAX8_01, PAX8_B, PAX_Q6
Rel/ankyrin family	CREL_01, NFKAPPAB50_01, NFKAPPAB65_01, NFKAPPAB_01, NFKB_C, NFKB_Q6, NFKB_Q6_01, P50RELAP65_Q5_01, RELBP52_01
RFX family	RFX1_01, RFX1_02, RFX_Q6
Runt class	AML_Q6, PEBP_Q6
RXR-like family	COUPTF_Q6, COUP_01, COUP_DR1_Q6, EAR2_Q2, GCNF_01, HNF4ALPHA_Q6, HNF4_01, HNF4_01_B, HNF4_DR1_Q3, HNF4_Q6, HNF4_Q6_01, PNR_01, TR4_03, TR4_Q2
SMAD class	SMAD1_01, SMAD3_Q6, SMAD4_Q6, SMAD_Q6, SMAD_Q6_01
SOX class	SOX2_Q6, SOX5_01, SOX9_B1, SOX9_Q4, SOX_Q6, SRY_02
SRF family	SRF_01, SRF_02, SRF_03, SRF_C, SRF_Q4, SRF_Q5_01, SRF_Q5_02, SRF_Q6
STAT class	STAT1STAT1_Q3, STAT1_01, STAT1_05, STAT1_Q6, STAT3STAT3_Q3, STAT3_01, STAT3_02, STAT3_03, STAT4_Q4, STAT5A_01, STAT5A_02, STAT5B_01, STAT_01, STAT_Q6
Steroid Hormone Receptors family	AR_01, AR_02, AR_03, AR_04, AR_Q2, AR_Q6, ERALPHA_01, ERR1_Q2, ERR1_Q3, ERR2_01, ER_Q6, ER_Q6_02, GR_01, GR_Q6, PR_01, PR_02, PR_Q2
TATA/TBP class	TATA_01, TATA_C, TRF1_01
T-Box class	BRACH_01, TBR2_01, TBX15_01, TBX15_02, TBX18_01, TBX22_01, TBX5_01, TBX5_02, TBX5_Q5
Thyroid Hormone Receptor-like family	FXR_IR1_Q6, FXR_Q3, LXR_DR4_Q3, LXR_Q3, PPARA_01, PPARA_02, PPARG_01, PPARG_02, PPARG_03, PPAR_DR1_Q2, PXR_Q2, RORA1_01, RORA2_01, RORA_Q4, T3R_Q6, VDRRXR_01, VDR_Q3, VDR_Q6

Supplementary Table 10. Replicate data quality and reproducibility. Each row represents a cell type for which two replicates were sequenced to comparable depth. Data quality scores for each replicate (columns two and three) are computed as the percentage of all reads that fall in DNaseI hotspots called on a 5 million tag subsample of each replicate. Column four is the correlation of the tag densities (150 bp sliding window tag count, stepping every 20bp) between the two replicates across chromosome 19.

Cell Type	Quality 1	Quality 2	Correlation
A549 (Human Lung Carcinoma Epithelial Cells)	0.4376	0.4086	0.9486
AG04449 (Fetal Buttock/Thigh Fibroblast)	0.4617	0.3886	0.9649
AG04450 (Fetal Lung Fibroblast)	0.4644	0.4019	0.9829
AG09309 (Adult Toe Fibroblast)	0.6948	0.4092	0.9388
AG09319 (Adult Gingival Fibroblast)	0.6695	0.4703	0.9895
AG10803 (Adult Abdomen Fibroblast)	0.7472	0.7097	0.9867
AoAF (Normal Human Aortic Adventitial Fibroblast Cells)	0.7162	0.6824	0.9892
BE2_C (Human Brain Neuroblastoma Cells)	0.6139	0.5567	0.9803
BJ (Normal Human BJ Skin Fibroblasts)	0.7488	0.5	0.9223
CACO2 (Colorectal adenocarcinoma)	0.7072	0.5	0.8989
CD20 (Human Leukapheresis Product)	0.5716	0.4473	0.898
GM04503D (Adherent Fibroblast Strain)	0.6456	0.6228	0.9839
GM04504A (Adherent Fibroblast Strain)	0.7513	0.7315	0.9532
GM06990 (GM06990)	0.5463	0.5463	0.9794
GM12865 (Female B-Lymphocyte Utah Pedigree 1459 Repository Linkage Family)	0.525	0.5036	0.9865
GM12878 (Lymphoblastoid cells)	0.5	0.4428	0.8361
H7_hESC_T14 (H7 human embryonic stem cells T14)	0.372	0.3622	0.984
H7_hESC_T5 (H7 human embryonic stem cells T5)	0.3431	0.3778	0.8399
HAc (Human Astrocytes-cerebellar)	0.4222	0.4152	0.9578
HAEpiC (Human amniotic epithelial cells)	0.7644	0.7512	0.9917

Cell Type	Quality 1	Quality 2	Correlation
HAh (Human Astrocytes-hippocampal)	0.4846	0.3093	0.9031
HAsp (Human Astrocytes-spinal cord)	0.4255	0.3919	0.9433
HBMEC (Human Brain Microvascular Endothelial Cells)	0.5433	0.417	0.9793
HBVSMC (Human Brain Vascular Smooth Muscle Cells)	0.3551	0.394	0.9489
HCF (Human Cardiac Fibroblasts)	0.688	0.608	0.9945
HCFaa (Human Cardiac Fibroblasts-Adult Atrial)	0.5183	0.4809	0.9679
HCM (Human cardiac myocytes)	0.7207	0.5102	0.9845
HConF (Human Conjunctival Fibroblasts)	0.5061	0.4838	0.9883
HCPEpiC (Human Choroid Plexus Epithelial Cells)	0.7418	0.6027	0.9854
HCT116 (Human Colorectal Carcinoma Cells)	0.4545	0.4015	0.9889
HEEpiC (Human esophageal epithelial cells)	0.5693	0.5493	0.9719
Hela (Cervical carcinoma)	0.5787	0.5816	0.9389
HepG2 (HepG2)	0.57	0.55	0.9168
hESCT0 (H7 undifferentiated human embryonic stem cells)	0.6353	0.5687	0.9698
HFF (Human Foreskin Fibroblast Cells)	0.5451	0.5395	0.9751
HFF_MyC (Human Foreskin Fibroblast Cells Expressing Canine cMyc)	0.4844	0.428	0.9579
HGF (Human Gingival Fibroblasts)	0.4832	0.4821	0.9799
HIPEpiC (Human iris pigment epithelial cells)	0.5596	0.54	0.9837
HL60 (Acute promyelocytic leukemia)	0.5888	0.5883	0.98
HMEC (Human Mammary Epithelial Cells)	0.4255	0.435	0.7662
HMF (Human Mammary Fibroblasts)	0.7977	0.7499	0.9814
HMVEC_dAd (Normal Adult Human Microvascular Endothelial Cells, Dermal-Derived)	0.3765	0.2111	0.9266

Cell Type	Quality 1	Quality 2	Correlation
HMVEC_dBlAd (Normal Adult Human Blood Microvascular Endothelial Cells, Dermal-Derived)	0.726	0.7035	0.9876
HMVEC_dBlNeo (Normal Neonatal Human Blood Microvascular Endothelial Cells, Dermal-Derived)	0.5289	0.4914	0.9701
HMVEC_dLyAd (Normal Adult Human Lymphatic Microvascular Endothelial Cells, Dermal-Derived)	0.5754	0.6281	0.9883
HMVEC_dLyNeo (Normal Neonatal Human Lymphatic Microvascular Endothelial Cells, Dermal-Derived)	0.5785	0.5394	0.9937
HMVEC_dNeo (Normal Neonatal Human Microvascular Endothelial Cells (Single Donor), Dermal-Derived)	0.5856	0.4094	0.9864
HMVEC_LBl (Normal Human Blood Microvascular Endothelial Cells, Lung-Derived)	0.4847	0.4701	0.9603
HMVEC_LLy (Normal Human Lymphatic Microvascular Endothelial Cells, Lung-Derived)	0.6046	0.5515	0.9903
HNPCEpiC (Human Non-pigment Ciliary Epithelial Cells)	0.6053	0.4433	0.9417
HPAF (Human Pulmonary Artery Fibroblasts)	0.7156	0.704	0.994
HPdLF (Normal Human Periodontal Ligament Fibroblast Cells)	0.6862	0.6071	0.9874
HPF (Human Pulmonary Fibroblasts)	0.6722	0.5917	0.977
HRCE (Human renal cortical epithelial cells)	0.6573	0.613	0.9817
HRE (Human renal epithelial cells)	0.534	0.43	0.9729
HRGEC (Human Renal Glomerular Endothelial Cells)	0.4239	0.3626	0.8689
HRPEpiC (Human retinal pigment epithelial cells)	0.7414	0.5997	0.982
HSMM (Normal Human Skeletal Muscle Myoblasts)	0.6368	0.5802	0.9205
HSMM_D (Primary Muscle myoblasts and myotubes)	0.4979	0.6013	0.8633
HUVEC (Primary Human Umbilical Vein Endothelial Cells)	0.4012	0.3225	0.8831
HVMF (Human Villous Mesenchymal Fibroblast Cells)	0.5905	0.6069	0.922
Jurkat (Acute T Cell Leukemia Lymphocyte)	0.4966	0.3938	0.9108
K562 (Chronic myelogenous leukemia)	0.5415	0.5205	0.9846
LNCap (Prostate Carcinoma)	0.6198	0.5305	0.9747

Cell Type	Quality 1	Quality 2	Correlation
MCF7 (Mammary gland adenocarcinoma)	0.4373	0.4356	0.912
NB4 (Human Acute Promyelocytic Leukemia Cells)	0.531	0.4814	0.9836
NHA (Normal Human Astrocytes)	0.5615	0.5075	0.968
NHBE_RA (Normal Human Bronchial Epithelial Cells)	0.3443	0.375	0.937
NHDF_Ad (Adult Normal Human Dermal Fibroblasts)	0.8045	0.7754	0.9864
NHDF_Neo (Neonatal Human Dermal Fibroblasts)	0.6976	0.6705	0.9951
NHEK (Normal Human Epidermal Keratinocytes)	0.3573	0.3119	0.9414
NHLF (Normal Human Lung Fibroblast)	0.7064	0.5159	0.9808
NT2_D1 (Pluripotent human Testicular Embryonic Carcinoma Cell Line)	0.3505	0.3099	0.9425
PANC1 (Pancreatic Carcinoma)	0.4176	0.4106	0.9852
PrEC (Prostate Epithelial Cells)	0.3233	0.3087	0.9707
RPTEC (Renal Proximal Tubule Epithelial Cells)	0.4866	0.4764	0.9807
SAEC (Small Airway Epithelial Cell)	0.6197	0.4274	0.9503
SKMC (Skeletal Muscle Cells)	0.8007	0.746	0.9867
SKNSH (Neuroblastoma)	0.6218	0.4001	0.8561
SK_N_MC (Human Brain Neuroepithelioma Cells)	0.3534	0.3382	0.9815
T_47D (Mammary Ductal Carcinoma)	0.5837	0.5505	0.9949
WERI_Rb1 (Retinoblastoma)	0.5459	0.3906	0.8239
WI_38 (Retinoblastoma)	0.6998	0.5744	0.9805
WI_38_TAM (Retinoblastoma)	0.6215	0.4675	0.9716

Supplementary Methods

1.1 DNaseI and histone modification protocols

DNaseI assays were performed using two different protocols (Duke and UW) on a total of 125 cell-types (85 from UW and 54 from Duke, with 14 cell-types shared; see **Supplementary Table 1**). Both protocols involve treatment of intact nuclei with the small enzyme DNaseI which is able to penetrate the nuclear pore and cleave exposed DNA. In the Duke protocol[10, 11], DNA is isolated following lysis of nuclei, linkers added, and the library sequenced directly on an Illumina instrument. In the UW protocol[12], small (300-1000 bp) fragments are isolated from lysed nuclei following DNaseI treatment, linkers are added, and sequencing of the library is performed on an Illumina instrument.

For H3K4me3 ChIP-seq, cells were crosslinked with1% formaldehyde (Sigma) and sheared by Diagenode bioruptor. The antibody used in the ChIP assay was 9751 (Cell Signaling) for histone H3 tri-methyl lysine 4. The ChIP DNA was made into libraries based on the Illumina protocol, and the size-selected libraries were sequenced on an Illumina Genome Analyzer IIx.

Sequence reads were mapped using aligner Bowtie, allowing a maximum of two mismatches. Only reads mapping uniquely to the genome were utilized in the analysis. Mapping was to male or female versions, depending on cell type, of hg19/GRCh37, with random regions omitted.

UW samples were typically sequenced to a depth of 25-35 million tags per replicate. Two replicates were produced for each cell type, and we chose the top-quality replicate of each for all downstream analyses. All UW replicates are screened for quality by measuring the percent of their tags falling in hotspots genome-wide. A "top-quality replicate" is the replicate with the highest such score for the given cell type. UW replicates tend to be very reproducible, with two replicates' tag densities across chromosome 19, expressed as linear vectors, usually achieving correlations ≥0.9. **Supplementary Table 10** lists the quality scores and chr19 tag-density correlations for all DNaseI replicates obtained by UW.

The Duke data was more variable in the depth to which libraries were sequenced; consequently we combined all replicates for each cell type and subsampled to a depth of 30 million tags. This made the Duke data approximately match the UW datasets.

We then identified DNaseI hypersensitive regions of chromatin accessibility (hotspots) and more highly accessible DNaseI hypersensitive sites (DHSs, or peaks) within the hotspots, using the hotspot algorithm (John et. al., 2011), applied uniformly to datasets from both protocols.

Briefly, the hotspot algorithm is a scan statistic that uses the binomial distribution to gauge enrichment of tags based on a local background model estimated around every tag. General-sized regions of enrichment are identified as hotspots, and then 150bp peaks within hotspots are called by looking for local maxima in the tag density profile (sliding window tag count in 150bp windows, stepping every 20bp). Further stringencies are applied to the local maxima detection to prevent overcalling of spurious peaks. Hostpot also includes an FDR (false discovery rate) estimation procedure for thresholding hotspots and peaks, based on a simulation approach. Random reads are generated at the same sequencing depth as the target sample, hotspots are called on the simulated data, and the random and observed hotspots are compared via their z-scores (based on the binomial model) to estimate the FDR.

Using the above procedure, we identified DHSs at an FDR of 1%. For the 14 cell-types assayed by both UW and Duke, we consolidated the two peak sets by taking the union of peaks. For any two overlapping peaks, we retained the one with the higher z-score; we consolidated hotspots by simply merging the hotspot regions between the two datasets. See section 1.2 below for DHS dataset availability.

Hotspots and peaks were called in the same way on the H3K4me3 ChIP-seq datasets, with the exception that reads mapped to the same location in the genome are all retained for DNaseI analysis, whereas only one tag per location is retained for ChIP-seq analysis.

In addition to the 125 DNaseI data sets sequenced at the "normal" depth of 25-35 million reads, we also make use in the section "Transcription factor drivers of chromatin accessibility" of one of several data sets we have sequenced to much greater depth. These are DGF, or digital genomic footprinting datasets, which were processed identically to the normal depth datasets. The K562 DGF dataset was sequenced to a depth of ~115 million reads. For the analysis referred to above, we merged the hotspots from UW K562 DGF with the hotspots called on the full, combined K562 replicates from Duke (~38 million reads, after combining reads).

Dataset availability:

- Aligned reads in BAM format for all datasets can be downloaded from the ENCODE Data Coordination Center at UCSC (http://genome.ucsc.edu/ENCODE/downloads.html) under the links for sections entitled
    - Duke DNaseI HS
    - UW DNaseI HS
    - UW DNaseI DGF
    - UW Histone

1.2 DHS Master List and its annotation

The DHSs called on individual cell-types were consolidated into a master list of 2,890,742 unique, non-overlapping DHS positions by first merging the FDR 1% peaks across all cell-types. Then, for each resulting interval of merged sites, the DHS with the highest z-score was selected for the master list. Any DHSs overlapping the peaks selected for the master list were then discarded. The remaining DHSs were then merged and the process repeated until each original DHS was either in the master list, or discarded.

For the genic annotations in Fig. 1b, we used all available Gencode v7 annotations[13, 14], i.e., Basic, Comprehensive, PseudoGenes, 2-way PseudoGenes, and PolyA Transcripts. The promoter class counts, for each Gencode annotated TSS, the closest master list peak within 1 kb upstream of the TSS. The exon class covers any DHS not in the promoter class that overlaps a Gencode annotated "CDS" segment by at least 75 bp. The UTR class covers any DHS not in the promoter or exon class that overlaps a Gencode annotated "UTR" segment by at least 1 bp. For the intron class, we define introns as the set difference of all Gencode segments annotated as "gene" with all "CDS" segments. The intron class covers any DHS not in the previous categories that overlaps the introns by at least 1 bp.

Each master list DHS is annotated with the number of cell-types whose original DHSs overlap the master list DHS. This is called the cell-type number for that DHS. Plots in Fig. 1c (made using the R function "beanplot" from the "beanplot" package) summarize the distribution of cell-type numbers for various categories of DHS annotations. Repeat categories for the LINE, SINE, LTR, and DNA

repeat classes were taken from UCSC RepeatMasker track annotations. We required that 50% of an individual master list DHS be contained in a repeat element in order to belong to its category. See below for the annotations used for the miRNA TSS category, for which 405 master list DHSs were within 100 bp. The promoter category is as described above; the distal category refers to the intergenic DHSs (as defined in panel Fig. 1b) located at least 10 kb away from any TSS.

Dataset Availability:

- FDR 1% peaks by cell-type available at
  - ftp://ftp-private.ebi.ac.uk/byDataType/openchrom/jan2011/combined_peaks
  - Individual cell-type files end in *fdr0.01.merge.pks.bed and *fdr0.01.bed
- 125 cell-type master list available at
  - ftp://ftp-private.ebi.ac.uk/byDataType/openchrom/jan2011/combined_peaks/multi-tissue.master.ntypes.simple.hg19.bed

1.3 miRNAs

miRNA coordinates were downloaded from miRBase (version 10)[15] and used to map miRNAs to their genomic locations. We removed the following miRNAs that are considered dead in the current release (version 18) of miRBase: hsa-miR-801, hsa-miR-560, hsa-miR-565, hsa-miR-923, hsa-miR-220a, hsa-miR-220b, hsa-miR-220c and hsa-miR-453. We changed the names of the following miRNAs to their current names in miRBase (version 18): hsa-miR-128a to hsa-miR-128-1, hsa-miR-128b to hsa-miR-128-2, hsa-miR-320 to hsa-miR-320a, hsa-miR-208 to hsa-miR-208a, hsa-miR-513-5p-1 to hsa-miR-513a-5p-1, hsa-miR-513-3p-1 to hsa-miR-513a-3p-1, hsa-miR-513-5p-2 to hsa-miR-513a-5p-2 and hsa-miR-513-3p-2 to hsa-miR-513a-3p-2. Some miRNAs (e.g., let-7a-1, let-7a-2) are expressed from multiple genomic locations, and hence all of the genomic locations were used to predict Transcription Start Site (TSS). We also identified miRNA genomic clusters by merging all miRNAs into clusters if they mapped to the same strand of the chromosome and were less than 10 kb apart.

To assign a TSS for each miRNA locus, we used RefSeq[16], AceView[17], ESTs, and Eponine predictions[18] downloaded from the UCSC genome browser (hg 18 version of the genome assembly; see below)[19]. We first identified miRNAs that were located within and in the same orientation as RefSeq gene. The TSS for these miRNAs was assumed to be the same as for the host genes, as it has been shown that miRNAs within host genes are generally co-transcribed from a shared promoter[20, 21]. For miRNA genes that did not match to RefSeq, we used AceView, which provides comprehensive transcriptional evidence from full length cDNAs and ESTs. We next used predictions by Eponine and EST clones to define the TSS of the remaining miRNAs. To identify EST clones, if both 5' and 3' ESTs were available from the same clone and formed a transcript containing the miRNA, the miRNA was considered expressed by this transcript and its TSS was the 5' end of the EST. For the remaining miRNAs whose TSS could not be found by the above methods, the position 500 bp upstream of the miRNA was taken as the TSS.

In the case of miRNAs that lie in genomic clusters, the TSS of the most 5' miRNA was assigned to all miRNAs in the cluster, because such miRNAs are expressed as a single primary transcript from a shared promoter[22]. MicroRNAs in the same host gene were considered to be in the same cluster irrespective of their distance from each other. All TSS coordinates were converted from hg18 to hg19 using the UCSC LiftOver tool.

Dataset Availability:

- miRNA TSS available at
    - ftp://ftp-private.ebi.ac.uk/byDataType/openchrom/jan2011/mirna_tss

1.4 Analysis of Repeat-Masked DHSs

RepeatMasker data was downloaded from the hg19 rmsk table associated with the UCSC Genome Browser. Repeat-masked positions cover 1,446,390,049 bp of standard chromosomes 1–Y. 1,257,126,829 bp (86.9%) of these are uniquely mappable with 36-bp reads. Even though much of the genome is derived from repetitive elements, evolutionary divergence has resulted in sufficiently different sequences that most positions can have reads uniquely mapped.

There are 1395 distinct named repeats in 56 families in 21 repeat classes. Data was analysed by repeat family because this gives a granularity suitable for display. A number of the classes are structural classes rather than classes derived from transposable elements. Bedops utilities[23] were used to count the number of DHSs which were overlapped at least 50% by each repeat family. The DHSs in the master list of sites from 125 cell types/tissues were tested for overlap with repeat families. **Supplementary Table 2** shows overlap statistics for families of elements with at least 5000 overlapping DHSs. **Supplementary Table 3** shows DHSs overlapping repeat-masked elements which we tested and found to be enhancers in transient assays.

1.5 Cells, transient transfection assay and reporter luciferase activity assay

PCR-amplified fragments spanning DHSs were typically 300–500 bp and encompassed the entire 150-bp DHS peak. To the 5' end of the each primer pair we added an additional 15 bp of DNA sequence (upstream sequence 5' GCTAGCCTCGAGGATATC-3' and 5'-AGGCCAGATCTTGATATC-3' in order to directionally clone via the Infusion Cloning System (Clonetech, Mountain View, CA) into pGL4.10[luc2] (Promega, Madison, WI), a vector containing the firefly luciferase reporter gene. All recombinants were identified by PCR and sequences verified. DNA concentrations were determined with a fluorospectrometer (Nanodrop, Wilimington, DE) and diluted to a final concentration of 100 ng/□L for transfections.

We performed the transient transfection assays on K562 and HepG2 cell lines by seeding 50,000 to 100,000 cells with 100 ng of plasmid in a 96-well plate. Twenty-four hours after transfection, the cells were lysed and luciferase substrate was added following the manufacturer's protocol (Promega, Madison, WI). We measured firefly luciferase activity using a Berthold Centro XS3 LB960 luminometer (Berthold Technologies, Oak Ridge, TN).

2 Transcription factor drivers of chromatin accessibility

2.1 ChIP-seq signal processing

Raw sequencing tags (BAM format) from ChIP-seq experiments in K652 cells were downloaded from the ENCODE DCC. Sequencing tags from replicate experiments were merged and mapped to hg19 with BWA using default settings. Tag densities were calculated in 150-bp sliding windows every 20 bp over the entire genome and normalized to 10 million reads. Aggregate transcription factor occupancy was computed by summation of the normalized ChIP-seq densities for individual factors (n=42). The pair-wise Pearson correlation was computed between DNaseI accessibility and transcription factor occupancy in DNaseI peaks using normalized DNaseI and the aggregate ChIP-

seq density at DHS peaks. Cumulative Pearson correlations of DNaseI density and ChIP-seq densities were iteratively calculated for the entire chromosome 19 by the sequential addition of transcription factor ChIP-seq densities in the order specified (**Supplementary Fig. 6b**).

2.2 Determining relationships between sequence motifs and chromatin accessibility

To obtain the results shown in **Supplementary Fig. 6c**, occurrences of motifs from the TRANSFAC database[24] were identified by running FIMO on the GRCh37/hg19 reference sequence with a detection threshold of $P < 10^{-5}$. For each of the 125 DNaseI cell types we scored each motif's association with chromatin accessibility by dividing the mean intensity (DNaseI tag count) of DHSs containing that motif by the mean intensity of all DHSs identified in that cell type. We then used the R package "beanplot" to visualize the distribution of this motif score across cell types.

2.3 ChIP-seq peaks and chromatin accessibility

ENCODE transcription factor ChIP-seq peaks for K562 were called using a uniform procedure as described[25], and downloaded from the ftp site below. The presence or absence of ChIP-seq peaks within accessible chromatin was determined by overlap or non-overlap, respectively, of each peak with deep-seq DNaseI hotspots in K562 (overlap by any amount was counted). Deep-seq K562 hotspots were constructed by merging hotspots for UW K562 DGF (sequenced at approximately 115 million reads) and hotspots for Duke K562 combined replicates (approximately 38 million reads). We used regular-depth K562 DNaseI tag density for the aggregate plots of **Supplementary Fig. 7a**.

Dataset Availability:

- Uniformly processed ChIP-seq peaks are available at
  - ftp://ftp-private.ebi.ac.uk/byDataType/peaks/jan2011/spp/optimal
- Deep-seq K562 hotspots are available at
  - ftp://ftp-private.ebi.ac.uk//byDataType/openchrom/jan2011/combined_hotspots/DGF

2.4 Quantification of the percentage of chromatin-bound protein

The percentage of total nuclear protein bound to chromatin was measured as described[26]. Briefly, K562 nuclei were isolated, as previously described[27], by resuspending cells at 2.5×10^6 cells/mL in 0.05% NP-40 (Roche) in Buffer A (15mM Tris pH 9.0, 15mM NaCl, 60mM KCl, 1mM EDTA pH 8.0, 0.5mM EGTA pH 8.0, 0.5mM Spermidine). After an 8-minute incubation on ice, nuclei were pelleted at 400*g* for 7 minutes and washed once with Buffer A. Nuclei were then transferred to a 37°C water bath and resuspended at 1.25×10^7 nuclei/mL in Isotonic Buffer (10mM Tris pH 8.0, 15mM NaCl, 60mM KCl, 6mM $CaCl_2$, 0.5mM Spermidine). After 3 minutes at 37°C, EDTA was added to a final concentration of 15mM and the sample was transferred to ice. The soluble and insoluble fractions were separated by centrifugation at 400*g* for 7 minutes. The total amount of nuclear protein that remained bound within the nuclei after this Isotonic Buffer wash was quantified using quantitative targeted proteomics as previously described[28].

2.5 Quantification of the percentage of nuclear protein present within heterochromatin

The percentage of total nuclear protein present within heterochromatin was quantified as described in[26]. Briefly, K562 nuclei were isolated, as previously described[27], by resuspending cells at 2.5×10^6 cells/mL in 0.05% NP-40 (Roche) in Buffer A (15mM Tris pH 9.0, 15mM NaCl, 60mM KCl,

1mM EDTA pH 8.0, 0.5mM EGTA pH 8.0, 0.5mM Spermidine). After an 8-minute incubation on ice, nuclei were pelleted at 400*g* for 7 minutes and washed once with Buffer A. Nuclei were then transferred to a 37°C water bath and resuspended at 1.25×10^7 nuclei/mL in MNase Buffer (25 U/mL MNase [Worthington], 10mM Tris pH 7.5, 10mM NaCl, 1mM $CaCl_2$, 3mM $MgCl_2$, 0.5mM Spermidine). After 3 minutes at 37°C, EDTA was added to a final concentration of 15mM and the sample was transferred to ice. The soluble and insoluble fractions were separated by centrifugation at 400 rcf for 7 minutes. The pellet was resuspended in 80mM Buffer B (10mM Tris pH 8.0, 80mM NaCl, 1.5mM EDTA pH 8.0, 0.5mM Spermidine), incubated at 4°C for 1 hour while rocking and then centrifuged at 2000 rcf for 8 minutes. The pellet was then washed sequentially for 1 hour each with 150mM Buffer B, 350mM Buffer B and 600mM Buffer B in a similar manner as the 80mM Buffer B wash except that the concentration of NaCl in Buffer B was adjusted. All supernatant fractions were cleared by centrifugation at 10,000 rcf for 10 minutes and any insoluble material was discarded. As previously described[29], the 350mM and 600mM solubilized fractions from MNase treated nuclei correspond to the heterochromatin fraction. The total amount of nuclear protein present within the 350mM and 600mM solubilized fractions was quantified using quantitative targeted proteomics as previously described[28]. To calculate the percentage of chromatin bound protein present within heterochromatin, for each factor the total amount of nuclear protein present within heterochromatin was divided by the total amount of that protein bound to chromatin.

3 Promoter DHS identification scheme

Our promoter DHS identification scheme consists of a joint analysis of DNaseI and H3K4me3 data. We focused our analysis on 56 cell-types for which we had joint data for both DNaseI and H3K4me3. The bulk of these cell-types were only studied by UW. For consistency we therefore restricted our analysis to UW datasets, even on those cell-types for which Duke and UW DNaseI data were both available. These 56 cell-types are indicated in **Supplementary Table 1**. The promoter identification scheme proceeds as follows.

For a given cell-type, we compute the 20th percentile D of the mean H3K4me3 density over a 550 bp window around Gencode v7 promoters overlapping a DHS from that cell-type. Within the set of promoters overlapping DHSs at the 20th percentile or greater for mean H3K4me3 signal, we look at the ratio of the H3K4me3 signal flanking the DHS to the signal at the DHS. More specifically, for each selected promoter, we compute the mean H3K4me3 signal over the 150 bp promoter DHS; over the 200 bp window immediately to the left of the DHS; and over the 200 bp immediately to the right of the DHS. For each flank we then compute the ratio of the flanking mean to the DHS mean, and retain the greater of these two ratios. We then find the 20th percentile across all selected promoters of these maximum ratios, R. To identify the "promoter DHS" from the pool of all DHSs within the given cell-type, we next find all DHSs that have mean 550 bp windowed (centered on the DHS) H3K4me3 density ≥ D. Within that set of DHSs, we flag all those that have ratio $R' \geq R$, where R' is the greater of the ratios of the mean H3K4me3 density in either of the flanking 200bp windows to the mean H3K4me3 density over the DHS. Note that the flanking window that gives the greater ratio also gives the prediction of the direction of the promoter.

We generated a set of 113,615 unique, non-overlapping promoter predictions across 56 cell-types as follows. First, all predictions for a given cell-type were partitioned into known-proximal and novel subsets. Known-proximal are all predictions within 1 kb upstream of annotated Gencode v7 TSS. Novel subsets are all remaining predictions, filtered so that no two novel predictions are within 5 kb of another prediction (novel or known-proximal), with preference given to predictions

with the greatest H3K4me3 flank ratio. Across cell-types, we generated a set of unique novel predictions by taking the union of all cell-type novel predictions and removing overlapping predictions, giving preference when there were overlaps to retaining the one with the greatest H3K4me3 flank ratio. This produced a total set of 44,853 unique novel predictions across cell-types. We generated an all-cell-types known-proximal list by taking all master-list DHSs that overlap any individual cell-type prediction that falls within 1 kb upstream of a Gencode annotated TSS, resulting in a total of 68,762 known-proximal positions, and a grand total of 113,615 unique, non-overlapping promoter predictions.

For the pie chart in Fig. 3c, Gencode coding and non-coding labels refer to the known-proximal predictions, with non-coding referring to any annotation with "RNA" in its biotype name, and coding referring to the remainder. The bar plot in the right portion of the panel further breaks down the novel predictions in terms of their supporting evidence by CAGE and EST annotations. For CAGE evidence we used a combination of Gencode and RIKEN cluster TSSs[14, 30]. RIKEN cluster TSSs were downloaded from the UCSC test browser. For a given cell type we used clusters for all cell localizations, using PolyA+ RNA. The overlaps shown here were relative to the pooling of RIKEN CAGE clusters for GM12878, K562, A549, Ag04450, H1Hesc, HelaS3, HepG2, and HUVEC cell types. Gencode CAGE cluster TSSs are made available through the ENCODE consortium[25]. Spliced ESTs were downloaded from the UCSC test browser. See **Supplementary Fig. 9** for the overlap of novel predictions with RIKEN and Gencode cluster TSS measured separately.

Overlaps with CAGE were tested for significance as follows. We focused on the 2279 K562 novel predictions, for which

973 (43%) are within 1 kb of a Gencode CAGE TSS
540 (24%) are within 100 bp of a Gencode CAGE TSS
2217 (97%) are within 1 kb of a RIKEN K562 CAGE tag
1987 (87%) are within 100 bp of a RIKEN K562 CAGE tag
1964 (86%) have a RIKEN K562 CAGE tag with the same orientation within 1 kb downstream
1590 (70%) have a RIKEN K562 CAGE tag with the same orientation within 100 bp downstream

There are 142,986 total K562 DHSs. Of these, we focused on the 93,672 of these that are not novel predictions, and not within 2500 bp of a known Gencode TSS. From this pool we chose random samples of size 2279; in addition, we randomly assigned a strand prediction to each sample element, in the same ratio of positive to negative orientations as assigned in the observed predictions (1149 positives, 1130 negatives). We generated 10,000 such samples, and none of them has the degree of overlap in any of the six measures above as those of the novel predictions, for a *P*-value less than 0.0001 for each result. The mean and standard deviation (SD) of the random sample results for each overlap are as follows:

within 1 kb of a Gencode CAGE TSS: mean = 65, SD = 8
within 100 bp of a Gencode CAGE TSS: mean = 23, SD = 5
within 1 kb of RIKEN K562 CAGE tag: mean = 1702, SD = 21
within 100 bp of RIKEN K562 CAGE tag: mean = 994, SD = 23
have a RIKEN K562 CAGE tag with the same orientation within 1 kb downstream: mean = 906, SD = 23
have a RIKEN K562 CAGE tag with the same orientation within 100 bp downstream: mean = 518, SD = 20

Dataset availability:

- Promoter predictions by cell-type, and unique novel and known predictions across cell-types available at
    - ftp://ftp-private.ebi.ac.uk/byDataType/openchrom/jan2011/promoter_predictions

4.1 RNA expression

For each cell line, total RNA was extracted in 2 replicates from 5×10^6 cells using Ribopure (Ambion) according to manufacturer's instructions. RNA quality was ascertained using RNA 6000 Nano Chips on a bioanalyzer (Agilent, Santa Clara, CA). Approximately 3 μg of total RNA for each sample was used for labeling and hybridization (University of Washington Center for Array Technology) to Affymetrix Human Exon 1.0 ST arrays (Affymetrix) using a standard protocol. Exon expression data were analysed through Affymetrix Expression Console using gene-level RMA summarization and sketch-quantile normalization method. Measurements from both replicates were then averaged. Raw data have been deposited in GEO under accession number GSE19090.

4.2 RRBS genome-wide methylation profiling

We downloaded RRBS methylation data for 19 cell lines from the "HAIB Methyl RRBS" track[31] of the UCSC Genome Browser. To measure methylation in each cell line, we combined counts for both strands in both replicates and removed CpGs with <8x coverage. We retained only CpGs monitored in at least 6 samples.

We applied a linear regression to measure whether methylation status is associated with accessibility. First, we generated a master list of DHSs found in any of the 19 cell lines. We then regressed accessibility onto the average proportion methylated of all monitored CpGs in a 150 bp region centered around the DNaseI peak. We tested only sites with both RRBS data for at least one CpG within the 150 bp window and ChIP-seq data for at least 6 cell lines. We excluded sites where the number of monitored CpGs differed by more than 4 among any two cell lines. We performed a linear regression at each remaining site, and used the R package qvalue to estimate a global FDR [32].

To assess the relationship between expression and TFBS methylation, we determined a set of putative binding sites for transcription factors, based on matches to database motifs inside DHSs where methylation was significantly associated with accessibility (see **Supplementary Table 5** for the mapping we used from TRANSFAC motif names to gene names). For each transcription factor, we regressed the average methylation at all of these motif instances onto the gene expression in each immortal cell type. We tested only motif models including a CpG.

5.1 Connectivity between promoter DHSs and distal DHSs

For the analyses described in section "A genome-wide map of distal DHS-to-promoter connectivity," we collapsed the DNaseI tag densities from 79 diverse cell types into aggregate densities within 32 categories of biologically similar cell types (**Supplementary Table 6**), and called consensus DHSs from these densities. We chose the 32 categories by hierarchically clustering the genomewide "present/absent" binary DHS vectors for the 79 cell types. For this part of our study, we defined a promoter DHS to be the consensus DHS overlapping a gene's TSS or nearest its TSS in the 5' direction. We identified 69,965 distinct promoter DHSs across the human genome, using the collection of TSSs in Gencode. A vector of aggregate DNaseI tag densities within each of the 32 categories was created for each promoter DHS. Similarly, we constructed 32-element tag-density

vectors for each of 1,454,901 consensus non-promoter DHSs located within 500 kb of a promoter DHS. We define a promoter/distal DHS pair to be "connected" if the Pearson correlation coefficient between the DHSs' tag-density vectors is 0.7 or higher. Where indicated, we used a correlation threshold of 0.8 for some analyses within this section. **Supplementary Table 7** contains the full set of promoter/distal DHS pairs connected at correlation threshold 0.7.

We compared the observed distribution of correlations with that of a null model in which we chose two DHSs at random that lie on different chromosomes, shuffled their cell-type category labels, computed their correlation, and repeated this 1,500,000 times. Using this null, we estimated the probability of observing a correlation >0.7 due to random chance alone to be 0.0102. We observed 1,454,901 non-promoter DHSs that were each within 500 kb of at least one of 69,965 promoter DHSs; we computed a total of 42,874,775 correlations for all such promoter/distal DHS pairs, and observed 1,595,025 of them to exceed 0.7, for an empirical probability of 0.0372 of observing a correlation >0.7, more than three times the probability within the null model. Using a binomial, we estimated the *P*-value for observing 1,595,025 or more correlations >0.7 out of 42,874,775, under this null, to be less than 10^{-100}. These 1.6 million high correlations were distributed among 578,905 distinct distal DHSs. The null model also shows that the promoters have more putative regulatory inputs than would be expected by random-chance assignments. Each promoter was found to be correlated with an average of 22.8 distal DHSs, with 84% of promoters correlated with multiple DHSs. The null model predicts an average of only 6.2 correlated DHSs per promoter, with only 67% of promoters correlated with two or more DHSs

5.2 Analysis of 5C and ChIA-PET data

For the analysis referenced in Fig. 5a, 5C[33] sequence reads were mapped to forward-reverse fragment pairs; raw data for only the highest read count interactions is displayed. Four enhancer sites match strong DHSs in the PAH region. We tested the three intronic DHSs shown in Fig. 5a by cloning these into pGL4.10[luc2], with the PAH promoter driving luciferase expression. We found each of these three DHSs stimulated PAH expression over twofold compared to the promoter-only construct. The site upstream of the promoter lies within the promoter HindIII fragment, and thus was not tested in our 5C experiments; however, this DHS has previously been implicated as an enhancer of PAH activity (see **Supplementary Table 4** for source).

FDR 1% peak interactions have been identified in several segments from the ENCODE pilot regions[6]. We used the subset of 5C peak interactions from K562 which contained at least one K562 DHS in the reverse (non-promoter) restriction fragment to obtain a distribution of maximal correlation scores for peak interactions; we assigned each peak interaction the highest correlation score observed within all promoter/distal DHS pairs in which the promoter DHS overlapped the forward fragment and the distal DHS overlapped the reverse fragment. We compared this distribution of scores to that of the highest-scoring DHS pairs for an interaction distance-matched control fragment for each of the peaks by applying a one-sided Mann-Whitney test to the medians of the distributions (**Supplementary Fig. 14b**).

The set of interactions detected via ChIA-PET in K562 cells in an earlier study[7] was filtered for interactions in which each tag overlapped a K562 DHS after padding by 100 bp on either side of the tag start. Correlation scores for interactions in which the ChIA-PET tags were at least 10 kb apart were tabulated. A control set was created by using the same distance distribution as the K562 ChIA-PET set and associating each original promoter site with a new simulated DHS. The set of correlation scores for the genome was filtered and, if a correlation score for the distance had been observed, it was added to the control distribution. The shuffling was repeated until the control set

had the same number of observations as the experimental set. The distributions were compared using a one-sided Mann-Whitney test (**Supplementary Fig. 14c**).

5.3 Gene ontology analysis of DHSs

To perform the analysis referenced in **Supplementary Fig. 14d**, we ranked all Gencode genes in descending order by the number of distal DHSs within ±500kb correlated with their promoter DHSs at a threshold of 0.7; for genes with multiple TSSs implicating multiple distinct promoter DHSs, we chose the promoter DHS with the highest number of connected distal DHSs. We used the rank-ordered list as input for a gene ontology analysis using GOrilla[34]; the search terms we used are listed in **Supplementary Table 8**.

5.4 Analysis of sequence motif pairs co-occurring in promoters and connected DHSs

We used FIMO[35] to identify all TRANSFAC motifs present in DHSs at confidence level $P < 10^{-5}$. We took the collection of all promoter DHSs across the genome, and for each one, recorded (1) the number of distinct motifs detected within it, (2) which motifs, if any, these were, and (3) the number of non-promoter DHSs within 500kb achieving correlation ≥ 0.8 with it. We then took the collection of all non-promoter DHSs across the genome, which tend to be narrower than promoter DHSs, and for each one, recorded (1) and (2). Together, these enabled us to create random promoter/distal motif pairs matched to the observed data.

Simulating random, matched motif data.
Specifically, we recorded the asymmetric square matrix (732 motifs × 732 motifs) of observed promoter/distal motif co-occurrence counts, and created two identically-sized matricies, each initialized to all zeroes. For each promoter DHS p containing m_p motifs and connected to d_p DHSs with correlation ≥ 0.8, we sampled (without replacement) m_p motifs from the observed distribution of motifs in promoter DHSs, and took d_p independent samples (with replacement) from the observed distribution of the number of motifs per distal DHS. (m_p and d_p were sometimes zero.) Then for each of the d_p numbers drawn, we sampled that number of motifs from the observed distribution of motifs in distal DHSs. (Each of the d_p independent samples was performed without replacement; replacement was allowed across independent samples. Some of the d_p sample sizes were zero.) All pairwise co-occurrences within the collections of sampled promoter motifs and distal motifs were tallied, while retaining the promoter and distal labels, and these tallies were added to the matrix of simulated random observations. After the tallies of random motif co-occurrences were accumulated within the random-matched matrix for all promoter DHSs, we compared each observed co-occurrence count with each random-matched co-occurrence count, and added 1 to the corresponding cell in the third matrix whenever the random-matched co-occurrence count was at least as large as the observed one. After performing one replicate randomization, this third, "tally" matrix consisted entirely of zeroes and ones.

P-*value estimation for co-occurrences of motifs and families of related motifs.*
We repeated this full procedure 100,000 times, which gave us a tally matrix whose tallies for specific motif co-occurrences ranged from 0 to 100,000. From this, we obtained an empirical *P*-value for each observed motif co-occurrence (i.e., for each nonzero element of the observation matrix) as the corresponding tally matrix element divided by 100,000. After obtaining *P*-values for co-occurrences of specific TRANSFAC motifs such as GKLF_02 within promoter DHSs and

USF_Q6_01 within distal DHSs, we investigated whether various groupings of specific motifs co-occur significantly often. We explored grouping motifs by their "pre-underscore strings," e.g., pooling BCL6_01, BCL6_02, BCL6_Q3 into "BCL6," and grouping them into families and classes defined by the structures of their associated proteins, e.g., pooling AFP1_Q6 and HOMEZ_01 into the "homeo domain with zinc-finger motif" family, or pooling HOX-like, NK-like, TALE-type and other homeo-domain factor families into the "homeo domain" class. (The family and class definitions we used, given in **Supplementary Table 9**, were adapted from http://www.edgar-wingender.de/huTF_classification.html, a web page actively maintained by Prof. Edgar Wingender, a co-founder and current board member of BIOBASE GmbH, which maintains the TRANSFAC database.) To compute empirical *P*-values for groupings of specific motifs, we randomly sampled specific motifs as described above, but summed the observed and random motif co-occurrences within our groupings of the specific motifs (e.g., any of BCL6_01, BCL6_02, BCL6_Q3 within a distal DHS co-occurring with either of AFP1_Q6 and HOMEZ_01 within a promoter DHS), and for each group × group co-occurrence, we estimated its *P*-value as the number of replicate data sets in which at least as many co-occurrences were present in the random matched data as in the observed data, divided by the number of replicates. **Supplementary Fig. 15b-c** illustrates enrichment of co-occurrences within 42 families and classes of motifs. The *P*-value matrix is clearly not symmetric (**Supplementary Fig. 15b**). Reassuringly and interestingly, closely-related motif families cluster together by membership in promoter DHSs (matrix rows, **Supplementary Fig. 15c**).

6.1 DNaseI pattern matching

For each cell type, a tag density file was prepared representing DNaseI cut counts observed in 150-bp windows shifted every 20 bp. Datasets were not normalized but represented similar levels of DNaseI sequencing. Summing these across all cell types, local maxima were identified and formed the universe of genomic locations subject to pattern search. For a given examplar region, all sites were ranked by a scoring function comparing the vector of DNaseI tag density to that of the exemplar site. The best matches were defined as those with the lowest sum of squared absolute differences in tag counts for each cell type between the two locations. Three representative patterns and the top 30 ranked pattern matches for two of them are shown in **Supplementary Figs. 16, 17**. When finding sites to be assayed in one or more particular cell types, a weight vector was applied to multiply all tag counts from those cell types by a small factor to increase the relative stringency of the match for those cell types.

6.2 Self-organizing map

In order to characterize the patterns of hypersensitivity across the 125 cell types of Supplementary Table 1, we constructed a self-organizing map (SOM) of the DHS data. We built a matrix of hypersensitivity scores from the maximum DNase-seq signal for each peak and cell type, resulting in a peak-by-cell-type matrix of DHS scores. We quantile-normalized the scores by cell type and then capped them at the 99th quantile (by setting the top 1% of scores to a maximum value), and then row-scaled the scores to a decimal between 0 and 1. After normalization, capping, and scaling, we built an SOM using the kohonen package in R. The SOM is an unsupervised clustering method that learns common DHS profiles in the data. Each node is initialized with a random DHS profile across cell types, and nodes are then iteratively adjusted according to the DHS profile of each peak. The SOM eventually assigns each peak to the node with the most similar hypersensitivity profile. Our SOM uses a hexagonal 35×35 grid (for 1225 total nodes). Because the software was unable to

handle all the data, we used a random sample of about 288,000 hypersensitive sites, reasoning that this would capture the major patterns.

To create the greyscale plot of **Supplementary Fig. 18c** showing the number of "strongly open" cell types, we set an arbitrary threshold (0.4) and counted cell types above this threshold. For the colour plot of **Supplementary Fig. 18a**, we assigned a colour to each cell type (**Supplementary Fig. 19**), and then assigned a colour to each node by taking a weighted combination of colours of cell types considered open in that node.

7 Measurement of nucleotide heterozygosity and estimation of mutation rate

We downloaded publicly-available genome-wide variant data for 54 individuals with no known familial relationships between them from Complete Genomics (ftp://ftp2.completegenomics.com/Public_Genome_Summary_Analysis/Complete_Public_Genomes_54genomes_VQHIGH_VCF.txt.bz2, Complete Genomics assembly software version 2.0.0). We validated the unrelatedness of the individuals using KING[36], a robust software package for inferring kinship coefficients from high-throughput genotype data. Two Maasai individuals in the dataset (NA21732 and NA21737) were not reported as related, but were found with KING to be either siblings or parent-child. We therefore removed NA21737 from the analysis, leaving us with genotype data from 53 unrelated individuals, with Coriell IDs HG00731, HG00732, NA06985, NA06994, NA07357, NA10851, NA12004, NA12889, NA12890, NA12891, NA12892, NA18501, NA18502, NA18504, NA18505, NA18508, NA18517, NA18526, NA18537, NA18555, NA18558, NA18940, NA18942, NA18947, NA18956, NA19017, NA19020, NA19025, NA19026, NA19129, NA19238, NA19239, NA19648, NA19649, NA19669, NA19670, NA19700, NA19701, NA19703, NA19704, NA19735, NA19834, NA20502, NA20509, NA20510, NA20511, NA20845, NA20846, NA20847, NA20850, NA21732, NA21733, NA21767. We filtered the variant sites to obtain only those for which full genotype calls were made for at least 20% of the individuals, treating partial calls (e.g. a genotype of A and N) as non-calls. From this filtered set, after first removing from consideration all sites within Gencode exons[13] and RepeatMasker regions (downloaded from the UCSC Genome Browser), we estimated allele frequencies for the locations of all variant sites occurring within the 53 genomes. For each variant with minor allele frequency p, the nucleotide heterozygosity at that site is $\pi = 2p(1 - p)$.

We computed the mean π per site within the DHSs of each of 97 cell lines by summing π for all variants within the DHSs and dividing by the total number of bases belonging to the DHSs, since $\pi = 0$ at invariant sites. To compare mean π per site between DHSs and fourfold-degenerate exonic sites, we used NCBI-called reading frames, summed π for all variants within the non-RepeatMasked fourfold-degenerate sites (thanks to Ian Stanaway), and divided by the number of sites considered. We estimated 95% confidence intervals on π per fourfold-degenerate site by performing 10,000 bootstrap samples.

To estimate relative mutation rates within the DHSs of each cell line, we downloaded human/chimpanzee alignments from the UCSC Genome Browser (reference versions hg19 and panTro2, http://hgdownload.cse.ucsc.edu/goldenPath/hg19/vsPanTro2/syntenicNet/), choosing the more conservative syntenicNet alignments; details can be found in http://hgdownload.cse.ucsc.edu/goldenPath/hg19/vsPanTro2/README.txt. Within the DHSs called in each cell line, we extracted the number of nucleotide differences between chimpanzee and human (d) and the number of bases aligned (n). We then estimated DHS-specific relative mutation rates μ per site per generation as $\mu = (d / n) / (2 \times 6 \text{ my} / 25 \text{ years/generation})$, with 6 million years being the approximate age of the human/chimp divergence[37].

Supplementary References

1. Bonauer, A., Boon, R. A. & Dimmeler, S. Vascular microRNAs. Curr Drug Targets 11, 943-9 (2010).
2. Townley-Tilson, W. H., Callis, T. E. & Wang, D. MicroRNAs 1, 133, and 206: critical factors of skeletal and cardiac muscle development, function, and disease. Int J Biochem Cell Biol 42, 1252-5.
3. Blackledge, N. P. et al. CTCF mediates insulator function at the CFTR locus. Biochem J 408, 267-75 (2007).
4. Cleutjens, K. B. et al. An androgen response element in a far upstream enhancer region is essential for high, androgen-regulated activity of the prostate-specific antigen promoter. Mol Endocrinol 11, 148-61 (1997).
5. Balasubramani, A., Mukasa, R., Hatton, R. D. & Weaver, C. T. Regulation of the Ifng locus in the context of T-lineage specification and plasticity. Immunol Rev 238, 216-32 (2010).
6. Sanyal, A., Lajoie, B., Jain, G. & Dekker, J. The long-range interaction landscape of gene promoters. Nature In Press (2012).
7. Li, G. et al. Extensive promoter-centered chromatin interactions provide a topological basis for transcription regulation. Cell 148, 84-98 (2012).
8. Bao, S. et al. Stem cell-like glioma cells promote tumor angiogenesis through vascular endothelial growth factor. Cancer Res 66, 7843-8 (2006).
9. Bigner, S. H., Bullard, D. E., Pegram, C. N., Wikstrand, C. J. & Bigner, D. D. Relationship of in vitro morphologic and growth characteristics of established human glioma-derived cell lines to their tumorigenicity in athymic nude mice. J Neuropathol Exp Neurol 40, 390-409 (1981).
10. Boyle, A. P. et al. High-resolution mapping and characterization of open chromatin across the genome. Cell 132, 311-22 (2008).
11. Song, L. et al. Open chromatin defined by DNaseI and FAIRE identifies regulatory elements that shape cell-type identity. Genome Res 21, 1757-67 (2010).
12. John, S. et al. Chromatin accessibility pre-determines glucocorticoid receptor binding patterns. Nature Genetics 43, 264-268 (2011).
13. Djebali, S. et al. Landscape of transcription in human cell lines. Nature In Press (2012).
14. Harrow, J. et al. GENCODE: The reference human genome annotation for the ENCODE project. Genome Res In Press (2012).
15. Griffiths-Jones, S., Saini, H. K., van Dongen, S. & Enright, A. J. miRBase: tools for microRNA genomics. Nucleic Acids Res 36, D154-8 (2008).
16. Pruitt, K. D., Tatusova, T. & Maglott, D. R. NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. Nucleic Acids Res 33, D501-4 (2005).
17. Thierry-Mieg, D. & Thierry-Mieg, J. AceView: a comprehensive cDNA-supported gene and transcripts annotation. Genome Biol 7 Suppl 1, S12 1-14 (2006).
18. Down, T. A. & Hubbard, T. J. Computational detection and location of transcription start sites in mammalian genomic DNA. Genome Res 12, 458-61 (2002).
19. Rhead, B. et al. The UCSC Genome Browser database: update 2010. Nucleic Acids Res 38, D613-9 (2010).

20. Saini, H. K., Griffiths-Jones, S. & Enright, A. J. Genomic analysis of human microRNA transcripts. Proc Natl Acad Sci U S A 104, 17719-24 (2007).
21. Rodriguez, A., Griffiths-Jones, S., Ashurst, J. L. & Bradley, A. Identification of mammalian microRNA host genes and transcription units. Genome Res 14, 1902-10 (2004).
22. Baskerville, S. & Bartel, D. P. Microarray profiling of microRNAs reveals frequent coexpression with neighboring miRNAs and host genes. Rna 11, 241-7 (2005).
23. Neph, S. et al. BEDOPS: High performance genomic feature operations. Bioinformatics In Press (2012).
24. Matys, V. et al. TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes. Nucleic acids research 34, D108--D110 (2006).
25. The_ENCODE_Consortium. Integrative Analysis of the Human Genome. Nature In Press (2012).
26. Stergachis, A. B., Wang, H., Maurano, M. T., MacCoss, M. J. & Stamatoyannopoulos, J. A. Extensive compartmentalization of human transcription factors within functional chromatin niches. Submitted.
27. Dorschner, M. O. et al. High-throughput localization of functional elements by quantitative chromatin profiling. Nat Methods 1, 219-25 (2004).
28. Stergachis, A. B., Maclean, B., Lee, K., Stamatoyannopoulos, J. A. & Maccoss, M. J. Rapid empirical discovery of optimal peptides for targeted proteomics. Nat Methods 8, 1041-3 (2011).
29. Henikoff, S., Henikoff, J. G., Sakai, A., Loeb, G. B. & Ahmad, K. Genome-wide profiling of salt fractions maps physical properties of chromatin. Genome Research 19, 460-469 (2009).
30. Lassmann, T. et al. CAGE analysis of cell compartments specific coding and non-coding RNA. Genome Res In Press (2012).
31. Varley, K. E. et al. Genome-wide characterization of dynamic DNA methyation across diverse human cell lines and tissues. Nature In Press (2012).
32. Storey, J. D. & Tibshirani, R. Statistical significance for genomewide studies. Proc Natl Acad Sci U S A 100, 9440-5 (2003).
33. Dostie, J. et al. Chromosome Conformation Capture Carbon Copy (5C): a massively parallel solution for mapping interactions between genomic elements. Genome Res 16, 1299-309 (2006).
34. Eden, E., Navon, R., Steinfeld, I., Lipson, D. & Yakhini, Z. GOrilla: a tool for discovery and visualization of enriched GO terms in ranked gene lists. BMC Bioinformatics 10, 48 (2009).
35. Grant, C. E., Bailey, T. L. & Noble, W. S. FIMO: scanning for occurrences of a given motif. Bioinformatics 27, 1017-8 (2011).
36. Manichaikul, A. et al. Robust relationship inference in genome-wide association studies. Bioinformatics 26, 2867-73 (2010).
37. McVicker, G., Gordon, D., Davis, C. & Green, P. Widespread genomic signatures of natural selection in hominid evolution. PLoS Genet 5, e1000471 (2009).